TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Supersemantics for Knowledge Extraction

Philipp A. Blohm

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. D. Frischmann
Prüfer der Dissertation:	1. Univ.-Prof. Dr. H.-W. Mewes 2. Univ.-Prof. Dr. R. Zimmer (Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 16.10.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 16.01.2015 angenommen.

Abstract

In several areas of biomedical research, the publication rate exceeds the human reading speed. From this imbalance a fundamental need for technical assistence to deal with the flood of scientific findings arises. Text mining tries to tackle this problem by automatically extracting relevant knowledge from scientific articles by applying different natural language processing procedures. Current text mining tools, however, are far away from reaching the quality of information extraction shown by manual curators.

In order to reach its full potential, different methods from different linguistic and other adjacent mathematical and logical fields need to be integrated into a text mining analysis. In this thesis, the term Supersemantics is coined as an umbrella term to refer to approaches that work towards achieving this integration. Supersemantics connects approaches from morphology, syntax, semantics, pragmatics and corpus linguistics as well as approaches from logic and statistics. It bridges the boundaries between typical units of linguistics like words, sentences and texts as well as external knowledge and integrates the information from each of them for a better analysis.

Besides the definition and overview of this emerging field, this thesis presents a variety of algorithms that integrate information from different fields, linguistic levels and sources. A word sense disambiguation algorithm improves the interpretation of terms by considering contextual information from a sentence. Extracted biological events are enriched by contextual information from predicate-argument structures. Event extraction tasks are simplified by considering the structure of the used documents. Word senses are derived from statistical corpus information. And text mining results are integrated with external structured resources to improve the quality of gene set enrichment analyses. In addition to these modules of supersemantic information integration, the path to a full-fledged supersemantic system incorporating all of these relevant elements is outlined both conceptually and by the implementation of two prototypes.

The results obtained in this work show that supersemantic algorithms can improve the quality and extend the coverage of existing text mining tools. Furthermore, the developed prototypes hint towards a possible approach to realizing a supersemantic framework and already outperform some existing text mining systems like Excerbt on tasks like protein event extraction.

Zusammenfassung

In vielen Bereichen der biomedizinischen Wissenschaft übersteigt die Publikationsrate die menschliche Lesegeschwindigkeit. Aus diesem Ungleichgewicht ergibt sich eine fundamentale Notwendigkeit technische Hilfmittel zur Verarbeitung der Flut an wissenschaftlichen Erkenntnissen einzusetzen. Text Mining nimmt sich dieses Problems an, indem es mithilfe von Verfahren des Natural Language Processing automatisiert Wissen aus wissenschaftlichen Publikationen extrahiert. Die Qualität gegenwärtiger Text-Mining-Verfahren reicht jedoch noch nicht an die Qualität der Informationsextraktion manueller Annotatoren heran.

Um das volle Potential einer Text-Mining-Analyse auszuschöpfen bedarf es der Integration von Verfahren aus unterschiedlichen linguistischen und anderen verwandten mathematischen und logischen Bereichen. In dieser Arbeit wird der Begriff Supersemantiken als Sammelbegriff für Ansätze, die diese Integration vorantreiben, eingeführt. Supersemantiken verbinden Ansätze der Morphologie, Syntax, Semantik, Pragmatik und Korpuslinguistik, sowie Ansätzen aus der Logik und Statistik. Sie überbrücken die Grenzen zwischen typischen linguistischen Einheiten wie Wörtern, Sätzen und Texten, sowie externem Wissen und integrieren die Informationen für eine bessere Analyse.

Neben einer Definition und einer Übersicht über dieses neue Feld werden in dieser Arbeit verschiedene Algorithmen präsentiert, in denen die Integration von Informationen aus unterschiedlichen Disziplinen, Informationsquellen und linguistischen Leveln praktisch umgesetzt wurde. Ein Word-Sense-Disambiguation-Algorithmus verbessert die Interpretation von Begriffen durch die Berücksichtigung von Kontextinformationen eines Satzes. Extrahierte biologische Events werden mit Kontextinformationen aus Prädikat-Argument-Strukturen angereichert. Eventextraktionsverfahren werden durch die Berücksichtigung der Dokumentenstruktur vereinfacht. Wortbedeutungen werden von den statistischen Eigenschaften eines Korpuses abgeleitet. Und Text-Mining-Ergebnisse werden mit externen strukturierten Wissensressourcen integriert um die Qualität von Genexpressionsanalysen zu verbessern. Über diese supersemantischen Informationsintegrationsmodule hinaus wird zudem der Weg zu einem umfassenden supersemantischen System, das alle relevanten Analysen umfasst, sowohl konzeptionell als auch anhand von zwei prototypischen Implementationen aufgezeigt.

Die erzielten Ergebnisse zeigen, dass supersemantische Algorithmen die Qualität und Reichweite von Text-Mining-Tools verbessern können. Darüber hinaus deuten die entwickelten Prototypen mögliche Herangehensweisen zur Realisierung eines superemantischen Frameworks an und liefern in Bereichen

wie Protein-Event-Extraktion schon jetzt bessere Ergebnisse als einige etablierte Text-Mining-Systeme wie z.B. Excerbt.

Acknowledgements

I thank Prof. Mewes and Volker Stümpflen for giving me the opportunity to complete this thesis and for their support and guidance. In addition, I want to thank the Bachelor and Master students I supervised: Anita Winkler, Tim Jeske, Felix Sappelt, Jon-Magnus Meier, Maximilian Herzog and Tobias Lutzenberger. They contributed in many parts of this thesis. Furthermore, I want to thank my colleagues at the IBIS and at Clueda, especially Benedikt Wachinger, Sofiane Latreche, and Tobias Sander. I thank HELENA for funding my research, my collaboration partners especially Silke Meiners, Goar and Dmitrij Frishman and Andreas Ruepp for the fruitful cooperations, as well as my family and friends for their support. And finally, I want to thank Robert Strache and dedicate the following quotation to him:

> "What about me? You didn't thank me!"
> "You didn't do anything ..."
> "But I like to be thanked!"
> (Homer and Lisa Simpson in "Realty Bites")

Table of Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Systems Biology . . . 5
- 1.3 Text Mining . . . 8

2 Supersemantics **23**
- 2.1 Levels of Context . . . 24
- 2.2 Pragmatics . . . 26
- 2.3 Discourse Analysis and Text Linguistics . . . 29
- 2.4 Why Supersemantics? . . . 31
- 2.5 A Supersemantic Analysis . . . 34
- 2.6 Related Work . . . 37

3 Sentence Contextualization **39**
- 3.1 Ambiguity and the Need for Disambiguation . . . 39
- 3.2 Context Matters - The Case of Word Sense Disambiguation . . . 41
- 3.3 Approach . . . 41
- 3.4 Results . . . 45
- 3.5 Conclusion . . . 46
- 3.6 Related Work . . . 47

4 PAS Contextualization **49**
- 4.1 Predicate-Argument-Structures as Context . . . 49
- 4.2 Negative results . . . 50
- 4.3 Extraction of Non-interacting Protein Pairs . . . 51
- 4.4 Confidence Score . . . 52
- 4.5 Results . . . 53
- 4.6 Conclusion . . . 55
- 4.7 Related Work . . . 56

5 Section Contextualization **57**
- 5.1 Section Information . . . 57

5.2 Rare Diseases . . . 60
5.3 Single-slot symptom extraction for a decision support tool . . . 61
5.4 Results . . . 63
5.5 Conclusion . . . 66
5.6 Related Work . . . 67

6 Text Contextualization **69**
6.1 Text Information . . . 69
6.2 Constraint-based anaphora resolution . . . 71
6.3 Results . . . 73
6.4 Conclusion . . . 76
6.5 Related Work . . . 77

7 Corpus Contextualization **79**
7.1 Corpus Information . . . 79
7.2 N-gram Analysis . . . 81
7.3 Word Space Models . . . 81
7.4 Word Space Visualization of Text Mining Results . . . 83
7.5 Evaluation . . . 87
7.6 Conclusion . . . 90
7.7 Related Work . . . 91

8 Integration of External Knowledge **93**
8.1 Functional Analysis of Gene Lists . . . 93
8.2 Functional Analysis Using Text Mining . . . 96
8.3 GO Analysis . . . 99
8.4 Application: mRNA Blood Expression Patterns in Epilepsy Patients Study . . . 101
8.5 Conclusion . . . 104
8.6 Related Work . . . 105

9 Towards a Supersemantic Analysis I - Shallow SRL **107**
9.1 Excerbt Restrictions . . . 107
9.2 Shallow Semantic Role Labeling . . . 112
9.3 Approach . . . 113
9.4 Evaluation & Application . . . 118
9.5 Related Work . . . 119

10 Towards a Supersemantic Analysis II - IntegreSSA **121**
10.1 Integrated analysis . . . 122
10.2 Evaluation . . . 131
10.3 Patient Record Analysis . . . 137
10.4 Related Work . . . 141

11 Discussion **143**
11.1 Integrated systems . . . 143
11.2 Knowledge Representation . . . 146
11.3 Analysis efficiency . . . 150

11.4 Implications of literature-based science 152
11.5 Learning to read 154
11.6 Learning to talk 156

12 Conclusion & Outlook **159**

A Word Sense Disambiguation Evaluation **163**

B Modular implementation of functional analysis tool **167**

C Additional results of GO analysis **171**

D Additional results of functional analysis of epilepsy study **173**

E Measuring semantic entropy of texts **175**

F POS tag set used in German IntegreSSA **183**

G Additional example of chunking with German IntegreSSA **185**

List of Abbreviations **187**

List of Figures **191**

List of Tables **195**

Glossary **197**

Bibliography **201**

1

Introduction

1.1 Motivation

> *"They say that Samuel Taylor Coleridge was the last person to have read everything. By the time he died there were too many books, they suggest, for any single brain to engage with. 'They', as usual, are wrong. There were already millions of books in Europe by the year 1500, just half a century after the first printed page flew from the first press. To read a million books in a lifetime you would have to read 40 a day for 70 years. I couldn't even smoke half that many cigarettes for half as long before giving up and it takes a lot longer to read a book than to smoke a cigarette, let me tell you.*
>
> *Philosophers, wits, novelists, cooks, poets, essayists, herbalists, mathematicians, builders, poets and divines had poured out more thoughts in that first 50 years than had been committed to paper or vellum in the previous thousand. And the rate only continued to increase as it approached this century's dizzyingly insane levels of oversupply. With so much flowing from so many different human brains, who can be bothered to read it? Not I, sir and madam, not I."*
>
> Stephen Fry, QI Book of Advanced Banter

As described by Stephen Fry, the amount of literature that is published is ever increasing. And with it, the relative amount of what a single person is able to read decreases accordingly. Not only has it become impossible to have read everything, but it has also become impossible to have read everything about biology, neuroscience or most other fields. Even very specific topics can no longer be covered completely. This situation has led to disturbing circumstances in many fields of science. Researchers entering a scientific field today will never be able to build on all the experiences of previous researchers - at least not without technical help.

To exemplify this point, if a scientist nowadays starts to research Alzheimer's, he is confronted with a vast plethora of scientific publications about the topic. PubMed (Pubmed, 2014), the most comprehens-

ive search engine for biomedical publications, returns 98,353 documents (as of the beginning of 2014) when entering 'alzheimer's'. If such a Samuel Taylor Coleridge of Alzheimer's was to read all of these, he would have to invest years of his life. More precisely, at an average reading speed of 250 words per minute[1], an average paper length of 8,467 words[2], 229 workdays per year[3] and a workday of eight hours the scientist would have to read for over 30 years. Furthermore, in the 30 years of reading, many more publications would have been published on Alzheimer's. Assuming that the current publication rate (8,187 articles in December 2013 about Alzheimer's) would stay constant[4], then there would be 248,066 new articles published by the time the scientist would be done reading the articles published up to today. This in turn would take another 76.4 years to read. The point where a scientist could read everything in his discripline are long gone. This example showed that even the days where he could read everything about his own topic - leaving aside publications on methods and related topics - have passed as well. Nowadays, on topics like Alzheimer's, the rate of publication is higher than the average reading speed. From this a fundamental need for other ways of knowledge transfer arises.

Scientists typically tried to minimize the effort of extracting knowledge from publications by using different kinds of heuristics. Scientific publications are structured in a way that they provide a very brief summary in the form of an abstract, thus potentially avoiding the need to read the whole paper. Furthermore, one tends to choose the paper one reads intelligently. Factors that are commonly used to filter the publications that one reads are the impact factor and reputation of the authors and journals of the publication. Additionally, the use of keywords describing a publication as well as more advanced information retrieval techniques help to filter the vast amount of possible reading matters.

While such heuristics provide a way of dealing with the information flood, they can at best postpone the problem temporarily. With exponential growth rates of scientific publications (see Figure 1.1), it will become impossible to read every relevant abstract or to read through all relevant publications that are tagged with certain keywords. The consequence of this will be overlooked or forgotten knowledge that might be relevant to current problems. Examples of these already occur throughout all disciplines.

In computer science, such a disregarding of knowledge happened e.g. for a procedure called back-propagation. The backpropagation algorithm can be used to train artificial neural networks to solve non-linearly separable problems. Prior to the invention of this algorithm the simpler Perceptron learning algorithm was used. This algorithm could only solve linearly separable problems but failed at problems like the XOR-problem. The introduction of backpropagation by Paul Werbos (Werbos, 1974) in 1974 could thus have immensly increased the potential of neural nets. Unfortunately, his findings were largely ignored for around 12 years.

The fact that the backpropagation algorithm was overlooked had far-reaching consequences for the whole field. The shortcoming of neural networks to solve non-linear problems, had led to a harsh criticism of artificial neural nets in general by Minsky and Papert (Minsky and Papert, 1969) in their book Perceptrons in 1969, which resulted in a nearly complete halt of research in the field until the mid-80s. Even more, together with other setbacks in related domains the criticism of neural nets lead to a period called the "AI winter" , in which funding for the whole field of artificial intelligence was heavily cut and the reputation of the field worsened. It was only 12 years after Werbos' invention that

[1] The average reading speed of an adult is at around 200-250 words per minute
[2] I averaged the length of 20 randomly chosen papers
[3] The amount of workdays in Bavaria (249 in 2014) minus 20 days of holidays
[4] which it most probably does not, instead the last years showed that it tends to grow exponentially

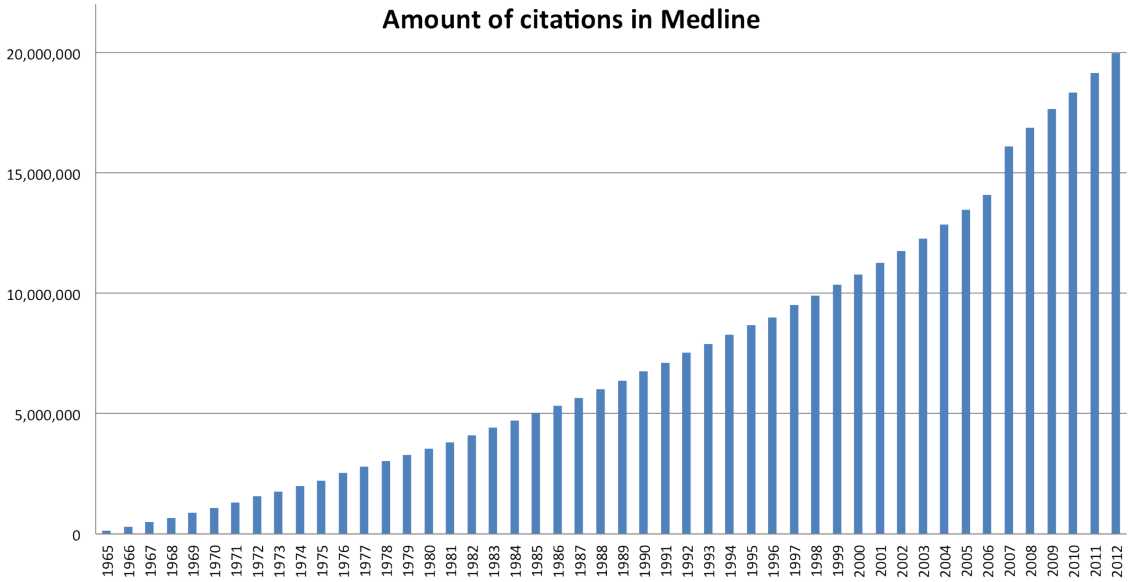


Figure 1.1: The diagram shows the growth of the Medical Literature Analysis and Retrieval System Online (MEDLINE), the main part of PubMed, the most comprehensive collection of biomedical knowledge, in the years between 1965 and 2012 (Pubmed, 2014).

the field was revived by Rumelhart, Hinton and Williams (Rumelhart et al., 1986) who rediscovered the backpropagation algorithm.

Another, more recent example from the biomedical domain was described by Steven A. Greenberg (Greenberg, 2009) in 2009. Greenberg analyzed how conflicting claims about the role of β amyloid inclusion body myositis patients were cited over time. β amyloid is a protein that accumulates in the brain of Alzheimer patients. Starting from four primary research papers that were supportive of the claim that β amyloid is produced by and injures skeletal muscle of patients with inclusion body myositis, and six critical ones, Greenberg analyzed which claims prevailed and how this happened. He found, that in the 12 years after the publication of the papers 94% of the citations were received by the supportive papers. Only 13 of the 214 citations were towards critical papers. Thus, in literature a very one-sided picture was drawn suggesting clarity about the claims. Even more, the existence of the critical results was largely forgotten.

Objectively, however, the critical papers were more trustworthy, since they were conducted from three independent labs, while the supportive ones all originated from the same lab. Furthermore, two of the critical papers were published by people from the very lab that published the supportive ones earlier on. Shortly after Greenbergs paper, Soscia et al. (Soscia et al., 2010) even published findings that β amyloid might actually be beneficial for Alzheimer patients. And Evans and Rzhetsky concluded that "the information cascade surrounding the deleterious effect of β-amyloid almost certainly prolonged experimental consideration of its possible beneficial role in Alzheimer disease and immunity" (Evans and Rzhetsky, 2011). Again important scientific findings were overlooked and again the scientific progress was delayed for over a decade.

While current heuristics are already error-prone, the challenges for scientific knowledge management are steadily increasing. In biology, the quite focussed approach on investigating the function of single genes is increasingly substituted by a systems approach that considers the interplay of many genes and environmental factors. With more complex models the required scientific background knowledge is increasing. In view of these ever more complex challenges, the typical heuristics are being stretched to their limits. The need for a technological way to support knowledge management arises. One field that tries to satify this need is text mining.

Most of the scientific knowledge that was gained over the years is published in scientific articles. These articles are aimed at a human audience, which is why they are written in normal continuous text, in text mining often referred to as unstructured text. In order to use the information given in the text, a computer program would, however, need the information in a structured form like e.g. a database or a graph. Only then it is possible to query such a program for typical scientific questions like “Which algorithm can train neural networks so that they solve non-linearly separable problems?” or “What is the role of β amyloid in inclusion body myositis patients?”. Thus, text mining employs methods from computer linguistics in order to extract structured information from unstructured text.

Text mining already succeeded at extracting mostly binary relations between proteins and other biological entities. Its error rate, however, is still clearly above that of human curators. The reason for this lies in the complexity of human language. Linguists struggled for decades to identify and solve typical error sources. Yet, there does not exist a system that could match the human ability to understand written texts. Possibly, the most central phenomenon responsible for this is the fact that every utterance needs to be interpreted within its context. Words, sentences and whole texts can vary in meaning depending on the circumstances within which they occur. Thus, including the context in a text mining analysis can prove to be a key to improving its performance.

Within this thesis ways to include context information at various stages in the text mining analysis are explored. This should on the one hand improve the precision of the analysis but also on the other hand extract additional contextual information from the text that current text mining systems often do not capture. The latter of these is a trend which occurred in the field of text mining in parallel to my work on this thesis. It is commonly referred to as contextualizaton. Since the core of this thesis is to introduce a framework that includes but goes beyond contextualization, a new term for these kind of analyses is coined here. Within the course of this work, such methods are referred to as supersemantics. This terminology should subsume all efforts to provide the principles, architectures and methods to produce better results and extract more information using context information.

The contribution made in this thesis is to define and give a comprehensive overview of supersemantics. Furthermore, it is to develop and apply methods that bridge levels of meanings and classical linguistic fields by incorporating contextual information. And finally, it is to provide an outlook into how a comprehensive supersemantic analysis incorporating all required aspects of contextualization might look like. In the remainder of this chapter the biological context of this work is described by giving an overview of systems biology. Additionally, text mining is described in order to provide the technical and scientific background of the work presented in this thesis.

1.2 Systems Biology

In 2003, the Human Genome Project announced the human genome sequence to be essentially deciphered. The "book of life" had been read the scientists of the project proclaimed (Noble, 2003). However, the so-called "blueprint of life" could not solve all the questions one hoped to find answers for. Ever since, a view has been established that the genome sequence only provides a "parts list" of life. But in order to fully understand how whole organisms work it needs to be understood how these parts interact. This view is the core idea behind systems biology (Newman, 2003).

Instead of investigating every component on its own the whole system is considered at once. Such a system is commonly represented as a network of interconnected elements which together form a whole. This network can be analyzed for its own properties - properties that are not necessarily displayed by any of its components on their own (Trewavas, 2006).

The observation, that a system can possess properties its components do not, is referred to by the term emergence. Emergence is one of the main motivations for systems biology. Only by considering the properties of the correct level of abstraction certain biological questions can be answered.

Emergent properties are attributed to a variety of phenomena, especially in complex systems. Frequently mentioned examples of emergence include behavioral patterns in swarms, consciousness and life itself (Bedau, 2003; Marsh, 2009). Swarm intelligence exceeds the abilities of its individuals. Consciousness and life appear as properties that are indivisible and not experienced by the single parts of the organisms that posses them. Each of the systems (a swarm, a brain, a body respectively) possessing such properties is made up of parts that by themselves do not posses the properties of the system or cannot account for them.

As can be seen in the simplified example from plant biology given in Figure 1.2 biological systems are systems of systems. Together they form a hierarchy of different levels of complexity with each level displaying its own emergent properties. Work in this field could also show that these levels interact with each other. Higher levels make use of the components of lower levels (downward causation) but also changes in lower levels, like mutations in DNA, can change higher level behavior (upward causation) (Trewavas, 2006).

With its perspective systems biology stands out against the traditional reductionist view of science. According to this paradigm complex problems are tackled with a "divide and conquer"-strategy. By dividing the problem into smaller parts, one retrieves a set of problems that are simpler to solve. E.g. instead of investigating the whole organism only single genes and their correlation to a disease are analyzed. While this methodology has produced plenty of great advances throughout all scientific disciplines, it seems to fail to deliver answers for some very relevant questions of today (Ahn et al., 2006).

The introduction of systems biology into biology as alternative to pure reductionism constitutes a paradigm shift. The history of reductionism in science dates back to Rene Descartes who saw the world as a clockwork mechanism that could fully be explained by the analysis of its parts. Later, Newton expanded Descartes approach laying the foundation of classical physics (Mazzocchi, 2008). Based on this a mechanistic biology evolved that saw organisms in analogy to a clockwork as deterministic

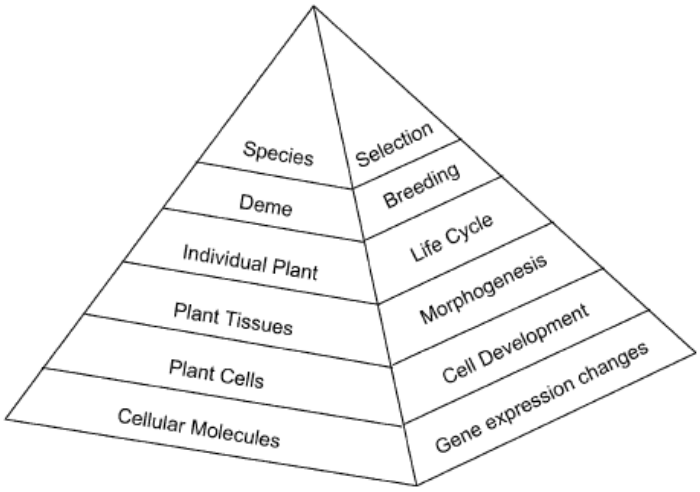


Figure 1.2: Hierarchy of the different levels of biological complexity in the domain of plant biology. Figure taken from (Trewavas, 2006).

machines. Loeb manifested this view in 1912 by stating that all animals of the same species would be behaving identical and predetermined like complex machines (Trewavas, 2006). In the 1950s a lot of researchers came from physics to molecular biology and further implemented the approach in the context of organisms (Mazzocchi, 2008). Accordingly, Francis Crick stated the goal of biology as "to explain all biology in terms of physics and chemistry" (Crick, 1966).

In line with these early reductionist ideas of scientific inquiry are experiments like knock-outs to determine the functional role of a gene. One focuses only on the gene, ignoring other causes or interrelations in the explanation of a certain phenotype. Such methods, however, have only restricted explanatory power when a phenomenon is a function of a complex network or can overemphasize the importance of a specific gene (Mazzocchi, 2008).

Correspondingly, in some areas of modern biology the limits of this approach seem to be reached. Diseases like cancer might be too complex to overcome them with the traditional strategies. Likewise in the neuronal system reductionism seems not to deliver enough insights for a comprehensive understanding of the subject (Mazzocchi, 2008). Instead, complexity as an important source for emergence needs to be considered in the biological study of higher level phenomenons like complex diseases.

While the idea of thinking of systems as a whole rather than a mere assembly of its sub-units dates back to Aristotle's famous quotation "the whole is something over and above its parts and not just the sum of them all", within the biological field the earliest roots of modern systems biology can be attributed to Jan Smuts (Smuts, 1926) who first introduced the term holism to the scientific community. Smuts pointed to emergent properties to argue against the reductionist approach. Furthermore, work in neurology and animal development made obvious that the behavior of individual parts of a system is dependent on the structure of the whole. This kind of orchestration of a system can be seen in situations of compensation, e.g. when hormone sensitivity increases as a result of low levels of hormones. The whole causally acts on its parts that cannot be explained by simply looking at the individual parts and trying to take the

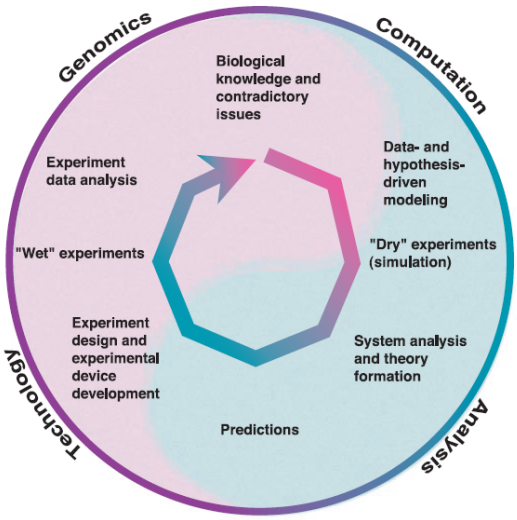


Figure 1.3: Hypothesis-driven loop of systems biology. Figure taken from (Kitano, 2002).

sum of them (Trewavas, 2006). Starting from these early observations, the investigation of complexity entered more and more in the biological worldview.

While complexity theory is a science on its own and does not have a single definition, complexity is often associated with some common properties. Complex systems are nonlinear, have build-in feedback methods and can produce spontaneous order (Ladyman et al., 2011). Systems biology is trying to account for this by modeling such systems in the form of networks and calculate stable states within these models. Furthermore, complex systems often inherit a certain robustness without a central control, have emergent properties, possibly a hierarchical order and consist of many elements (Ladyman et al., 2011).

The understanding of a biological system, according to Kitano (Kitano, 2002), can be achieved by investigations in four key properties: the system structures, the system dynamics, the control method and the design method. The system structure can be examined by the analysis of networks of interacting biological elements. Temporal analyses are necessary for understanding the dynamics of a system. Typical patterns that control the behavior of a cell or another biological unit can be identified and used in therapies. And finally, understanding the design principles of a biological system makes it possible to investigate biological processes by simulations and opens up the possibility to design desired phenotypes. Systems biology research focused in all of these areas may prove to greatly advance the field (Kitano, 2002).

The roots of this field are in the first half of the 20th century, but only with the introduction of methods in the field of molecular biology it became feasible to analyze systems systematically. With high-throughput measuring technologies and next-generation sequencers, biologists were finally equipped to look at biological processes on a higher level (Kitano, 2002).

As a new paradigm on doing research in biology, systems biology also stands for a new workflow of how to integrate the work of bioinformaticians and biologists. This so-called loop of systems biology

(see Figure 1.3) models the frequent exchange of information between the two fields. Starting with a biological question, bioinformaticians can build models on the existing data and derive new hypotheses from those models that behave correctly in their simulations. Biologists in turn can test these hypotheses and rule out the models that make false predictions, thus narrowing the space of possible models. By deriving the hypotheses exclusively on the basis of the original data the systems biology loop stands for a strongly hypothesis-driven approach to science (Kitano, 2002).

In order to build complex models, the bioinformatician needs to acquire the relevant data. While some biologists put great effort into creating databases collecting the results of the field, a lot of results are still scattered across the many publications of the different journals and not recorded in a database. To use this hidden information and to create tools that give scientists a better overview over their field, automated methods extracting information from publications are needed. The field of text mining is trying to provide these methods.

1.3 Text Mining

For ages, humans have been collecting their knowledge of the world within books. Libraries grew and in times of the information age masses of information became available to everyone with an internet connection. However, availability does not necessarily imply findability. Already in the 17th century, with growing sizes of text collections, librarians saw the necessity for cataloguing systems to facilitate the retrieval of relevant texts. From this need, in modern days the field of Information Retrieval (IR) originated which studies the ways text documents can be searched effectively in databases and other text collections to fulfill different information needs (Miner et al., 2012).

However, the recent information overload resulted in the demand for more sophisticated ways to collect information. The traditional keyword-based information retrieval approaches did not suffice anymore. From the need for the advanced organization, maintenance and interpretation of information from textual sources the field of text mining arose (Ananiadou et al., 2006).

Definition

The most common definition of text mining dates from Marti Hearst's 1999 paper "Untangling Text Data Mining", in which he discriminated text mining, or text data mining how he called it, from bordering areas like information retrieval, text categorization and computational linguistics. According to Hearst text mining is the analysis of unstructured texts with the goal of uncovering new, previously unknown information (Hearst, 1999).

This new information can come in the form of associations, patterns or clusters of related texts. Different information pieces can be combined to synthesize new information. In the context of systems biology, text mining can be used to generate hypotheses that can afterwards be tested by experimental biologists (Ananiadou et al., 2006).

Biomedical text mining is a highly interdisciplinary field. As it is treated within this thesis, it is at the intersection between biology, linguistics, computer science, artificial intelligence and information

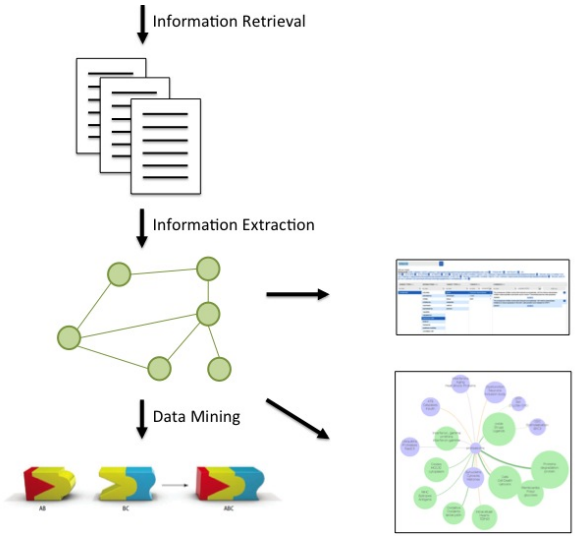


Figure 1.4: The workflow of a prototypical text mining system. Document are fist collected using information retrieval methods and then analyzed using information extraction methods. The structured information is then used to create new hypotheses or directly stored in data bases or used to build other tools. The depiction of the hypothesis generation step is taken from (Evans and Rzhetsky, 2010).

science. Biology asks the questions and provides important domain specific knowledge that is necessary to facilitate the analysis. Linguistics, or more specifically computer linguistics, provides the tools for analyzing the texts. Computer science solves the problems that arise from the need to deal with huge amounts of data in an efficient way. Information science offers ways of representing the information in a structured way. Finally, artificial intelligence is included. This area contributes in two ways: first it provides machine learning methods for generating new hypotheses and secondly inference algorithms might become a valuable part of future text mining tools when automated reasoning about the collected information might become more important (I will talk more about automated reasoning in sections 7.1 and 11.4).

While different applications might ask for deviating strategies, the general approach to text mining can be subdivided into the subtasks of information retrieval (IR) , information extraction (IE) and data mining (see Figure 1.4). In this workflow, information retrieval methods are used to collect the relevant text documents. Afterwards, information extraction methods are applied to extract facts from the documents. Finally, in the data mining stage new information is inferred on the basis of the extracted facts (Ananiadou et al., 2006). In many applications the results of the information extraction stage can also be used immediately for providing resources of structured data or tools helping scientists in their daily work without the application of data mining tools.

All three stages of this pipeline should be described in the following. Since both IR and IE heavily depend on algorithms borrowed from Natural Language Processing (NLP) the introduction of these stages is preceded by an overview over some of the relevant methods and resources from NLP.

Natural Language Processing

The main problem when trying to extract information from freely written text is the complex nature of modern languages. Firstly, languages are open systems. People can invent new words, talk about previously unseen objects and events and thereby extend or change the language. Furthermore, languages have complex grammars, in which words can be combined in different order or be inflected to change their meanings (Vogt, 2005).

Additionally, languages are designed to allow for ambiguities, which poses another fundamental problem for reliable information extraction from text. Multiple possible interpretations can occur on several levels. Terms can have multiple meanings (lexical ambiguity), sentences can have multiple grammatical structures resulting in different meanings (syntactic ambiguity) and sentences can be read in different ways (semantic and pragmatic ambiguity) (Ceccato et al., 2004).

The range of this problem might be illustrated by looking at systems that try to determine the grammatical structure of a sentence. In 1987, Martin et al. famously reported that their system assigned 455 possible structures to the sentence "List sales of the products produced in 1973 with the products produced in 1972." (Martin et al., 1987).

The problem broadens even when transfered to the biomedical domain. Different studies (Chen et al., 2005; Fundel and Zimmer, 2006; Tuason et al., 2004) showed the extraordinary degree of ambiguity in papers published in this domain. According to Chen et al. ambiguity among terms describing genes can reach up to 85% (Chen et al., 2005). Even worse, some of the ambigue terms are very common english words like ’we’, ’fold’, ’gel’ or ’inactive’ which frequently occur in biological papers (Fundel and Zimmer, 2006). As demonstrated in Figure 1.5 each term in the sentence ’Homer and Maggie go to Paris to meet a large, hairy clown and chat with the happy tinman.’ can also refer to a biological entity. This phenomenon is mainly due to the use of aliases instead of official names.

Since it is so complicated to deal with text, especially in the biomedical domain, text mining needs to borrow methods from experts that dealt with the pitfalls of modern languages for longer. Thus, a variety of methods developed in the field of natural language processing is used when trying to extract information from biomedical texts. This part of the thesis is meant to give an overview of the most relevant methods of NLP.

Natural language processing acts as a tool box for text mining in various ways. Firstly, within the NLP community several resources have been created that can be utilized when solving a language processing problem. Most notably different corpora have to be mentioned here. These corpora can act as gold standard to evaluate the performance of an algorithm for a given task and can be used to train machine learning algorithms for solving the problem. Within the biomedical domain the GENIA corpus (Ohta et al., 2002) and the data sets created within the BioCreAtIvE competitions (Arighi et al., 2011; Hirschman et al., 2005a; Krallinger et al., 2008) are among the most noteworthy. Furthermore, there exist word lists containing terms especially relevant for certain tasks and thesauri that list synonyms or semantically related terms. Finally, there are ontologies that model the concepts and their relationships within a specific domain. For genes and their products e.g. the Gene Ontology (Ashburner, 2000) was designed to create a standardized representation of the inter-relations of the concepts.

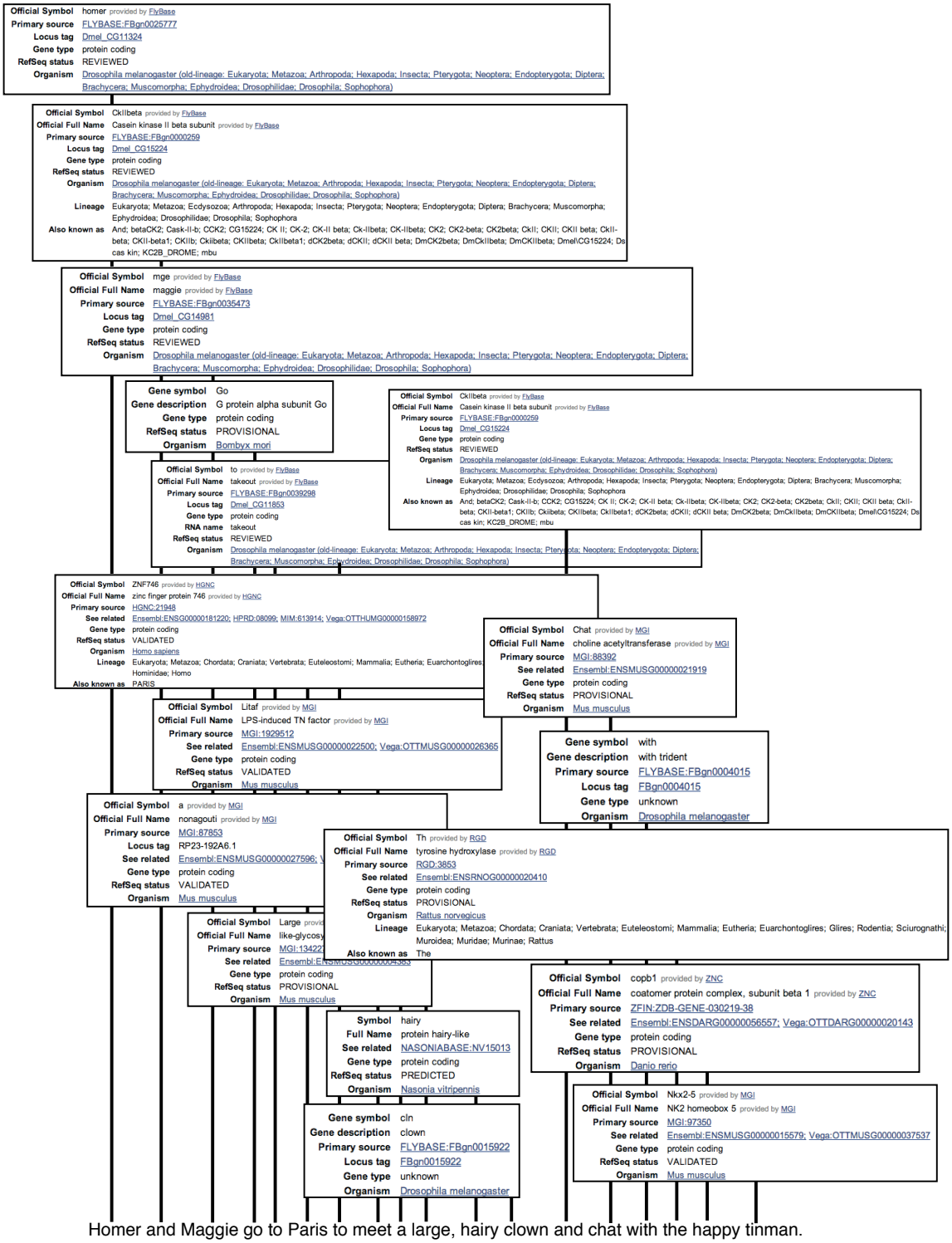


Figure 1.5: An example sentence to show the high level of ambiguity in gene names and aliases. All words in this sentence are names or aliases of genes or gene products.

Depending on the task, different representations of the given text are more useful. When one wants to know what topic a text is about e.g. it can be helpful to simply ignore the very common words in a text. Such a procedure is called stop word removal. On the other hand when one is interested in extracting more sophisticated information from a sentence it would be better to leave these words in, since a lot of stop words are function words that determine the structure of the sentence.

Because of such different requirements for different problems, a series of basic methods to change the representation of a text have been developed which can be applied as preprocessings. Besides stop word removal, these include tokenization (detecting term boundaries), sentence splitting (detecting sentence boundaries) and stemming (Cohen and Hunter, 2004). The idea of stemming is to change the declined form of a word to its word stem, e.g. the words 'stems' and 'stemming' would be stemmed to 'stem'. This is useful when all occurrences of a concept should be analyzed together no matter in which form they appear. While stemming only cuts off the ending of words to remain with the stem, lemmatization goes a step further and tries to map words to their base form (Ingason et al., 2008). Thus, the terms 'wife' and 'wives' would have the same lemma, but not the same stem.

These preprocessing steps can be sufficient, when the text should only be represented by the words it is made of. A well-established representation of this form is the so-called bag-of-words model. Here, after stop word removal and possible normalization (stemming or lemmatization) the collection of the remaining terms is considered the feature vector of the text. This methodology is widely applied in the field of information retrieval (Jiang et al., 2004).

In order to extract more detailed information from text, more sophisticated preprocessing procedures are needed. In this case, not simply the word but the complete sentence is taken into consideration. The surrounding of the word can be used to disambiguate the meaning of the word or structural information for the words or parts of the sentence can be determined.

Classically, computer linguists used constituency or dependency parsers that constructed parse-trees which represented the syntactical structure of sentences. Later on, also shallow parsers or chunkers were developed which tried to get as much structural information as possible out of the sentence without having to determine the overall grammatical structure. These methods cluster words into labeled 'chunks' that represent e.g. noun phrases or verb phrases. The motivation behind the development of these systems was the need to produce better performant systems for the application to very large collections of text (Delmonte, 2005).

Both full and shallow parsers commonly work on so-called part-of-speech (POS) tags. The task of POS tagging is to assign each word in a sentence its word class. These word classes are distinguished on the basis of the syntactic and morphological behavior of the words (Voutilainen, 2004). Methods for determining the right POS tag reach from rule-based methods over transformation-based learning (the system learns rules by itself) to Hidden Markov Models (Martinez, 2012).

Going beyond the syntactic analysis of sentences the field of Semantic Role Labeling (SRL) tries to tackle the semantic interdependencies of different parts of a sentence. The idea behind SRL is to answer the question "Who did What to Whom, and How, When and Where?" (Palmer et al., 2010). Thus, such systems assign roles to clusters of words which tag them as the agent, recipient, item and so forth of a relation commonly described by a verb.

There exist two major schemes to assign roles to parts of the sentence. One is defined in the FrameNet corpus (Baker et al., 1998), the other one in the PropBank corpus (Palmer et al., 2005). While the former has a very wide range of hierarchically arranged roles, the latter tries to generalize by combining roles to a more concise tagging scheme.

The majority of SRL systems are based on Machine Learning algorithms and are making use of a large variety of features. Here, the syntactic structure is commonly taken into account by deriving features from parse-trees generated by different parsers (Palmer et al., 2010). However, lately also systems avoiding parse-trees have been established. This approach is taken in order to create more performant systems that qualify for application in large text collections. Of these the Senna system (Collobert et al., 2011a) is the most noteworthy, reaching comparable results with the parser-based systems.

All of the methods described above can be made use of when designing a text mining system. Besides the already described tags and the bag-of-words model, some additional important forms of representation should be mentioned here. The vector space model (Salton et al., 1975) is often used in information retrieval systems. Here, whole documents are represented as high-dimensional vectors. Each entry in the vector corresponds to a particular term which does (non-zero value) or does not (zero value) occur in the document. By comparison of such vectors documents relevant to a query, which can also be represented as a vector, can be found.

The vector space model has recently also been extended to capture semantic representations between different text types. Besides measuring similarities between documents, they were also successfully applied to words by taking into consideration the context vector of the word and word pairs with the vectors representing their shared context (Turney and Pantel, 2010). Further representations of words can be produced by extracting word embeddings from trained language models or by using clustering methods (Turian et al., 2010b).

Information Retrieval

The term information retrieval (IR) was coined in 1950 by Calvin Mooers (Mooers, 1950) but the roots of the field go back way further to the origins of library systems. Nowadays, IR stands for every system that aims to find a relevant subset of informational items from a larger collection. This can involve finding relevant videos, images or, as relevant in the context of TM , texts. The relevance of IR increased immensely with the rise of the internet where also its most prominent applications can be found - the search engines (Larson, 2012).

The general workflow of an IR system is shown in Figure 1.6. Both the query and the document are first processed with NLP techniques like parsing and normalization procedures. Indexing procedures then turn the document and the query into their final representation (e.g. the vectors of the vector space model). In this form the two entities are compared and ranked according to their similarity. Based on this scores the relevant documents are finally returned (Larson, 2012).

Within the domain of biomedical TM the information retrieval stage can either be used to restrict the considered text collection to relevant documents in order to focus the process on a certain domain. Alternatively, it can also simply be reduced to collect as many documents from the biomedical domain as possible in order to widen the range of the output.

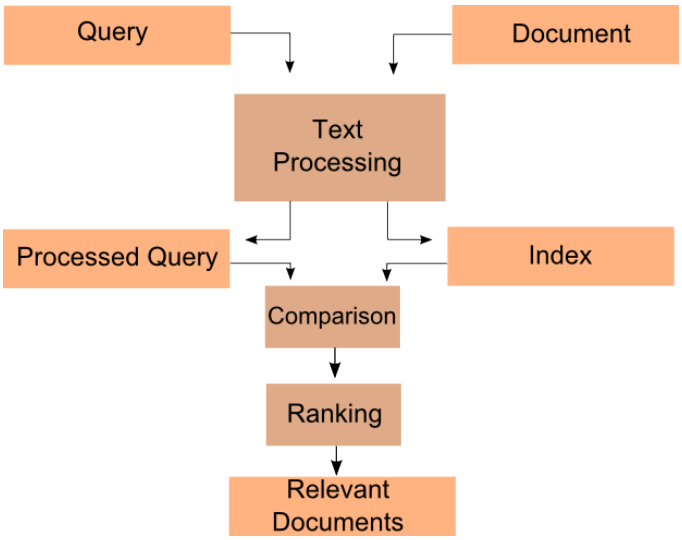


Figure 1.6: Overview of the components of an Information Retrieval system. Figure based on (Larson, 2012).

Furthermore, the process of acquiring relevant documents is closely related to information retrieval. Most biomedical text mining systems rely on comprehensive resources of digitally accessible publications like PubMed (Pubmed, 2014) or PMC (PMC, 2014). While such resources cover many publications at least partly, there are still many that cannot be accessed this way. The reasons for this can be that they are published in proprietary journals that do not allow such a distribution or simply by journals not covered by the resources. One solution to reach the latter of these are web crawlers. A web crawler is a computer program that starts with an initial set of websites and uses the hyperlinks in these find new interesting sites and documents. Then these are in turn analyzed for further hyperlinks, so that the crawler can expand in the search space (usually the internet). Web crawlers can be focussed on certain domains and types of documents (Chakrabarti et al., 1999). Thus, they can be used to find publications not covered by the resources used by one's text mining system.

Information Extraction

The objective of information extraction (IE) is to search through textual documents and to extract information that is relevant to a certain interest. This can concern the extraction of entities, relations or events (Hobbs and Riloff, 2010). In the case of the extraction of entities this process is called named entity recognition (NER) . In the case of relations its called relation recognition (RR) and event recognition (ER) when events are detected (Ananiadou et al., 2010).

Named entity recognition is the task of identifying proper names in a text. Traditionally, this approach is focused on detecting mentions of persons, locations and companies. However, depending on the domain of the IE system this can be extended or shifted to different entities (Hobbs and Riloff, 2010). Within the biomedical domain the detection of gene names or phenotypes are common applications for NER.

There are different strategies to extract named entities from texts. Dictionary-based approaches compare occurring terms to existing lists of terms for which the corresponding class is known. More sophisticated approaches to NER include rule-based, classification-based and sequence-based approaches. Rule-based systems decide on a specific role for a term on the basis of hand-crafted or syntactic rules (e.g. capital letters or digits can be hints for protein names). The remaining two strategies consider NER as a classification problem and employ machine learning algorithms to solve it. Here, the classification-based approaches classify token-wise (single words or phrases) while the sequence-based methods consider a whole sequence of tokens at once (Leser and Hakenberg, 2005).

Additionally, different strategies can be combined for hybrid approaches. The more sophisticated systems are based on NLP methods like POS-tagging and normalization. Furthermore, word sense disambiguation and abbreviation resolution can be applied, since ambiguities and varying notations are among the biggest difficulties of NER (Leser and Hakenberg, 2005).

A relation describes a connection that exists between several, usually two, terms. Thus, the most frequent representation of a relation is a pair of entities. A relation has a type, e.g. in the context of genes a regulation is a typical relation, and the two entities building up the relation have fixed RelationSemanticssemantic roles. In the regulation relation e.g. the entity with the semantic role of the agent is the one regulating the other one. The second entity has a role called theme and is the one being regulated (Ananiadou et al., 2010).

Events on the other hand always correspond to concrete incidents in the real world. They can be described by relations but can also have more participants and the semantic roles corresponding to the different participants vary depending on the specific event. Some of the participants might be necessary like the localized entity in a localization event, while others might only occur optionally in certain sentences. For example events can often be detected without knowing their cause. The cause can, however, be given as additional information. For the extraction of events a detailed analysis of the structure of the sentence has to be performed. Within the biomedical domain among other things event extractions can be used to identify protein-protein interactions, support pathway construction and improve the search abilities of text collections like MEDLINE (Ananiadou et al., 2010).

Approaches to the extraction of events or relations are diverse. The first systems relied on rather naive hand-crafted rules to detect the relations. An easy way to create an automated system is by predicting a relation between entities on the basis of co-occurrences of the entities. Furthermore, machine learning (ML) algorithms can be applied to the problem. ML methods can either be used to learn suitable rules or templates for the extraction, thus, avoiding the extensive effort of the manual creation of these rules. Alternatively, ML algorithms can be used to learn to tag a sentence according to the relations or events immediately, or to detected patterns in the text in an unsupervised fashion. Sophisticated linguistic rule systems can derive relations way more reliable than the naive rule systems. Apart from that, discourse-oriented approaches take a wider context of the sentence into account e.g. by the prior identification of event-related sentences. Finally, NLP-based approaches build on the structural analysis of the sentence to increase the performance of the prediction. This can be done by building on parsing or semantic role labeling (Cohen and Hersh, 2005; Hobbs and Riloff, 2010).

Since NER is part of the relation and event extraction in many of the approaches (e.g. co-occurrences and the NLP-based extraction) the preprocessing steps word sense disambiguation and acronym resolution can be applied here, too. Furthermore, anaphora resolution, the resolution of referential

Figure 1.7: Prototypical representation of an information extraction pipeline.

expressions like pronouns, can be applied to increase the recall of the methods. By replacing pronouns or other references by their respective named entity more relations can be found.

The different information extraction steps are commonly arranged in a pipeline of software modules. A prototypical representation of such a pipeline is given in Figure 1.7. First sentences are preprocessed by detecting sentence and token boundaries. The latter of which are POS tagged. This is the input of the sentence analysis that uses syntactic and/or semantic analyses. Finally, named entities are found within the syntactic/semantic roles. These patterns are then transformed in events where applicable. The details of the single steps can vary as described above. Furthermore, there exist various variations of this pipeline depending on the approach and purpose of the tool. For example co-occurrence analyses can omit the POS tagging and sentence analysis steps.

Data Mining

Data mining is defined by an objective rather than by a certain procedure. It is a broad collection of methods used to extract information, that is difficult to acquire, from any kind of data. This hidden information commonly can come in the form of patterns, that can be extracted e.g. as association rules, as clusters, that order the information according to its internal structure, or as classifications, that order the information in accordance with known examples (Coenen, 2011).

Many applications of data mining depend on the use of machine learning algorithms to classify or cluster the data. The most prominent here are support vector machines, artificial neural nets, random forests and various clustering methods. When applied in the context of text mining these can be used to classify texts. Furthermore, texts can be summarized or opinions can be mined making use of free texts or questionnaires (Coenen, 2011).

Within the biomedical domain additional applications and methods arise. The data mining stage of the text mining workflow is often used to produce new hypotheses. Famously, Don Swanson was the first to implement the ABC model for hypothesis generation from texts of different scientific fields. The simple idea was to make use of the principle of transitivity to correlate formerly unconnected concepts. If in one piece of scientific literature it was written that A causes B and in another that B causes C, then it could be concluded that A causes C (Swanson, 1986).

Making use of this simple principle Swanson produced the hypotheses that fish oil could moderate the symptoms of Raynauld's blood disorder (Swanson, 1986), that magnesium deficits are correlated to migraine (Swanson, 1988), which were both later confirmed experimentally, and that indomethacin was correlated to Alzheimer's disease (Swanson and Smalheiser, 1994).

Apart from that, data mining was already applied to several areas in bioinformatics like e.g. in sequence-based functional classification of proteins. Such approaches might be improved by the integration with text mining. Here, text mining can be used as a way to make unstructured data available. Thus, structured and unstructured information can be combined to deliver better results (Ananiadou et al., 2006).

If relations are extracted from text these can be composed to form graphs. On the basis of this, graph mining techniques (data minig on graphs) can be applied at this stage of the pipeline. A common technique here is to detect frequent or especially meaningful sub-graphs, so-called motifs. Furthermore, algorithms from the growing field of network analysis can be applied to find hidden information in the graph structure and to produce new hypotheses (Coenen, 2011).

Excerbt

The development of a comprehensive text mining system is a very complex endeavor that cannot be accomplished by a single person on its own. For this reason, it is important to describe the text mining infrastructure that existed when I started my thesis, respectively that was built in parallel to my work. A text mining system called Excerbt was developed between 2006 and 2012 at the Institute of Bioinformatics and Systems Biology at the Helmholtz Centre in Munich. Excerbt was first designed by Thorsten Barnickel (Barnickel, 2009) in his Phd thesis and developed further by Benedikt Wachinger (Wachinger, 2013) in his Phd thesis and Robert Strache (Strache, 2012) in his Master's Thesis. The system serves as basis for my work and hence should be introduced in this section.

In order to extract biological events from unstructured texts a semantic analysis of the text is needed. Most semantic analysis methods, however, depend on syntactic analyses that take long processing times. This poses a problem considering the current size of corpora like MEDLINE and the anticipated further growth of the collection of biomedical literature. In order to create a system that can efficiently be applied to large corpora and that can scale up in the future, Excerbt was based on an efficient semantic analysis called Senna (Collobert et al., 2011b) that omitted time-consuming syntactical analyses.

Senna is based on a deep neural net that performs several natural language processing tasks simultaneously. Most importantly it creates predicate-argument structures (PAS) from sentences. Such a structure is trying to capture all semantic aspects about a single proposition that are mentioned within one sentence. It formalizes the answers to the question "Who did What to Whom, and How, When, and Where?" by assigning different roles that represent the semantic dimensions of the proposition. An overview of the roles assigned by Senna is given in Table 1.1. Most importantly, the roles ARG0 and ARG1 correspond to the actor and theme of the PAS. The actor is the one that actively initiated something and the theme is the entity something is done to. For example in the sentence "Ice cream causes overweight." 'ice cream' would be the ARG0 and 'overweight' would be the ARG1.

Excerbt uses Senna to extract different biological events by searching the ARG0, ARG1 and Pred roles for biological entities from its internal ontology and verbs indicating biological events respectively. Its ontology is based on a multitude of common biological resources - e.g. gene and protein names are, among others, taken from Entrez Gene, Swissprot and Interprot. In Excerbt all abstracts from PubMed and all freely available full-text articles from PMC are analyzed.

Table 1.1: The roles assigned by the Senna role labeling tool. The roles correspond to the Propbank definition. The example sentences are based on or taken from the examples from the Propbank annotation guidelines Bonial et al. (2010).

Role	Name	Example
REL	Predicate	Mr. Bush **met** him privately.
ARG0	Argument 0	**Mr. Bush** met him privately.
ARG1	Argument 1	Mr. Bush met **him** privately.
ARG2	Argument 2	Mary left **her daughter** her pearls.
ARG3	Argument 3	They will remain **on the list**.
ARG4	Argument 4	Let them return **to the city**.
ARG-A	Secondary Agent	**John** walked his dog.
ARGM-COM	Comitatives	I sang a song **with my sister**.
ARGM-LOC	Locatives	Mr. Bush met him **in the White House**.
ARGM-DIR	Directional	I got kicked **out of the class**.
ARGM-GOL	Goal	The child fed the cat **for her mother**.
ARGM-MNR	Manner	Mr. Bush met him **privately**.
ARGM-TMP	Temporal	Mr. Bush met him **on Tuesday**.
ARGM-EXT	Extent	The shares closed **at $3.75, off 25 cents**.
ARGM-REC	Reciprocals	He would build it **himself**.
ARGM-PRD	Secondary Predication	He will join them **as a director**.
ARGM-PRP	Purpose Clause	They will return **for future meetings**.
ARGM-CAU	Cause Clause	They will stay **because of the interview**.
ARGM-DIS	Discourse	**But** for now, they stayed.
ARGM-MOD	Modals	John **ca**n't keep up with her.
ARGM-NEG	Negation	John ca**n't** keep up with her.
ARGM-LVB	Light Verb	Yesterday, Mary **made** an accusation.
ARGM-ADV	Adverbials	**Happily**, she sang.
ARGM-ADJ	Adjectivals	His **shocking** abuse outraged them.

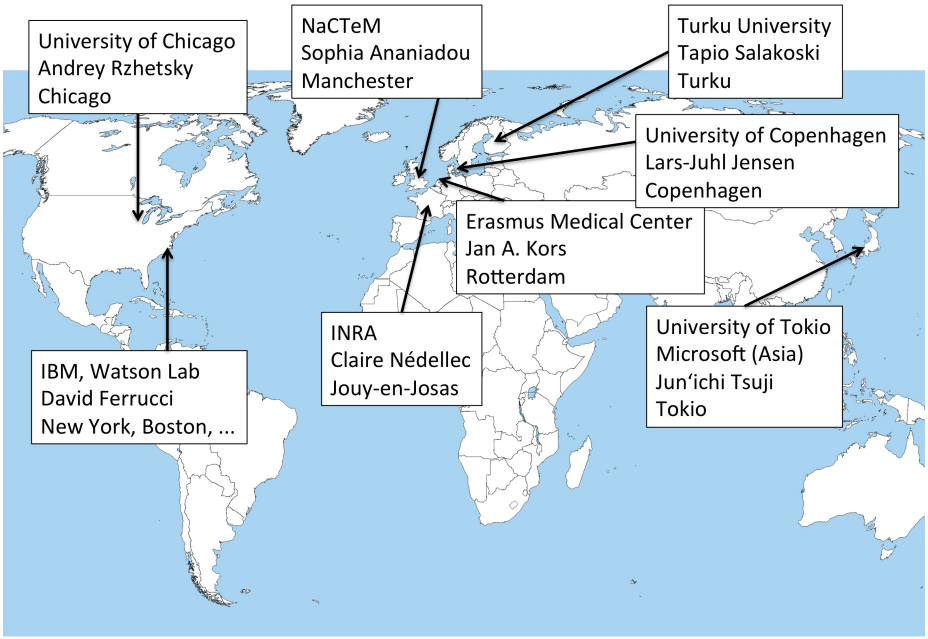


Figure 1.8: The text mining landscape. Some of the most influential research groups and researchers are shown.

Excerbt contains around 175 million different biological events extracted from 1.44 billion sentences. Thus, an event is backed by 8.2 sentences on average. Excerbt's ontology contains around 580,000 entities. There have been attempts to evaluate Excerbt on the BioNLP 2011 and the BioInfer corpus as well as by a comparison to SIDER database. The obtained results were mostly poor, however, it was argued that this might have been due to insufficiencies in the data sets (Strache, 2012; Wachinger, 2013).

Current Text Mining Landscape

In the years since Marti Hearst first introduced text mining, the field developed quickly and diversely. Today, many groups all over the world work on text mining and develop tools that may help scientists in their daily work. There are competitions held that should provide an objective quality assessment of the different systems and resources created to support and test systems. While the field is too large to cover comletely, at least the most influential elements with respect to the biomedical domain should be introduced here. This is intended to provide a comprehensive picture of the state-of-the-art in text mining and give an idea of the background in front of which this thesis was developed.

Figure 1.8 takes the text mining landscape literally and shows some of the most important text mining institutions. Most notably, the National Centre for Text Mining (NaCTeM) , the Watson Lab and the Turku University should be mentioned here. NaCTeM is the first national center for text minnig. It is located in Manchester and headed by Sophia Ananiadou. Researchers from the NaCTeM are among the most active members of the text mining community. They are involved in organizing competitions

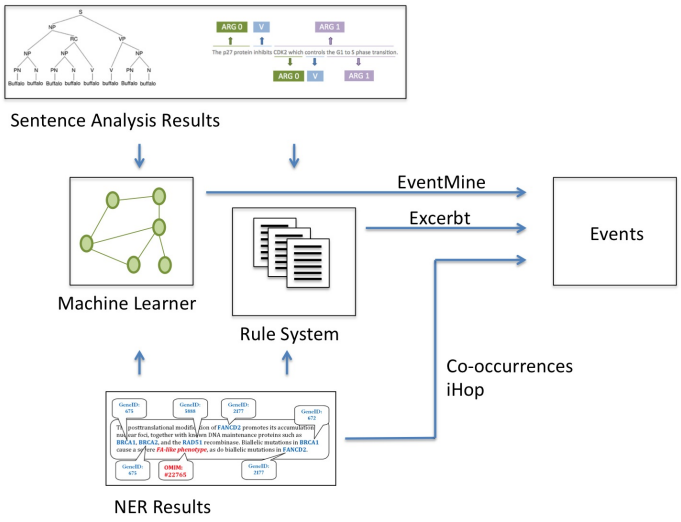


Figure 1.9: Simplified schematic description of different approaches to text mining.

like the BioNLP 2013 shared task (Nédellec et al., 2013), implementing a variety of text mining tools like BioContext (Gerner et al., 2012) and even the publication of a journal specifically dedicated to text mining (Journal of Biomedical Semantics). The Turku University is noteworthy because of its development of TEES (Björne et al., 2009). This text mining system can be considered the state-of-the-art in the domain of protein-protein event detection. It won several competitions and was integrated in different other text mining systems. Finally, the Watson Lab of IBM deserves mention. The researchers of the technology company developed a program called Watson (Ferrucci, 2011) that was able to compete in the television quiz Jeopardy. It performed so well that it was able to beat the best human players in a direct competition. Furthermore, this publicity stunt introduced text mining also to a considerable part of the non-scientific community.

A more detailed view of the text mining community is given in Figure 1.10. In line with typical representations of text mining results, a network seemed like the most appropriate way to display this. As one can see, there are complex involvements of different research groups with each other and on various topics. Besides the already mentioned Excerbt and BioContext, iHOP (Hoffmann and Valencia, 2005), EVEX (Van Landeghem et al., 2013), COREMINE (COREMINE, 2014) and String (Franceschini et al., 2013) are popular text mining resources that are available through web interfaces. As shown in Figure 1.9, the different text mining systems follow different strategies. Co-occurrence-based systems, like iHOP or String, immediately associate entities found by a NER system. In contrast to this, more sophisticated systems take results of linguistic parsers and/or semantic analysis tools into account. EventMine for example is based on the Enju (Sagae et al., 2007) and Gdep (Sagae and ichi Tsujii, 2007) parsers and Excerbt is based on Senna (Collobert et al., 2011b). The linguistic sentence analysis results are then combined either by rules (Excerbt) or by machine learning algorithms (EventMine, TEES) to detect biomedical events.

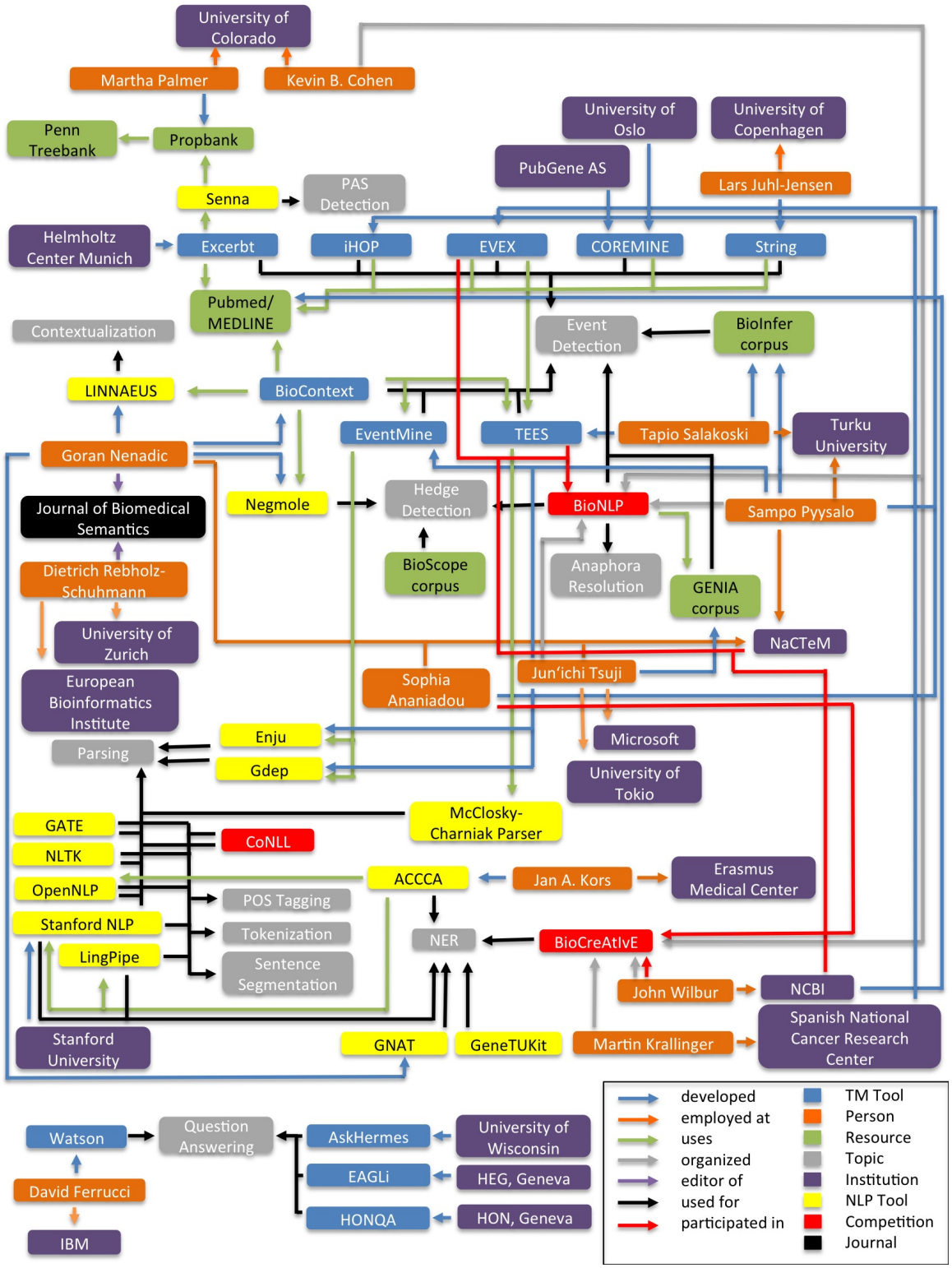


Figure 1.10: Network of some of the most influential entities of the biomedical text mining community.

While all of these systems have their main focus on event extraction, there exist several tools for more specific tasks: Negmole (Sarafraz and Nenadic, 2010) detects negations, LINNAEUS (Gerner et al., 2010) detects species names, and GNAT (Hakenberg et al., 2011) and GeneTUKit (Huang et al., 2011) detect proteins. Analogously, to the different tools there exist different annotated corpora that can be used for training and evaluation of systems. Most notably here, the BioInfer (Pyysalo et al., 2007a) and the GENIA (Kim et al., 2003) corpus can be used for event detection and the BioScope corpus (Szarvas et al., 2008) for hedge detection (negations and speculations). Since a vast majority of text mining tools depend on natural language processing methods, there is a close connection to the field of NLP. Some text mining researchers like Tsuji developed their own parsers while others drew on existing NLP frameworks like NLTK (Bird, 2006), OpenNLP (The Apache Software Foundation, 2010), Stanford NLP (University, 2011), GATE (Cunningham et al., 2011) or LingPipe (Carpenter and Baldwin, 2011) or combined existing ones into comprehensive frameworks (e.g. ACCCA (Kang et al., 2012)). The different foci are also represented by different competitions. While the BioNLP competitions focus most on event extraction, the BioCreAtIvE challenges (Hirschman et al., 2005b) traditionally are more concerned with named entity recognition and the CoNLL shared tasks (Ng et al., 2013) concentrate on natural language processing methods. The question answering domain is comparatively separated from the information extraction community. While Watson's focus is not on biology, there still exist biomedical question answering systems like AskHermes (Cao et al., 2011), EAGLi (Gobeill et al., 2009) and HONQA (Cruchet et al., 2009).

The overview, of course, only provides an extract of the domain. Many other active research groups and tools exist, which have not been covered here. Furthermore, because of the dense interconnections in large parts of the community possibly some connections have been missed in this overview picture. Other kinds of resources have been left out. For example, many biological ontologies and databases are frequently used as vocabularies for different text mining systems. And finally, many detailed applications or very specified tools have not been covered due to the sheer amount of them. To counteract this unavoidable shortcoming, each of the following chapters that describes a supersemantic application developed in this work is accompanied by a section that briefly summarizes the relevant related work for this specific topic.

2

Supersemantics

Biomedical text mining has changed over recent years. It developed from rather rudimentary approaches like co-occurrences over more sophisticated event extractions to the inclusion of contextual information. The trend of contextualization developed in parallel to the process of writing this thesis and is still ongoing. Supersemantics is a neologism coined in this thesis. As presented in this work, it encapsulates contextualizations and tries to provide a comprehensive framework for it. Furthermore, it tries to include important trends and design paradigms for successful future text mining. The idea that should be promoted in this thesis breaks with the traditional ways in which linguistic analyses are performed in the sense that it ignores and bridges the boarders between adjacent linguistic fields for the purpose of a more practical solution.

The elements of language convey meaning on different levels. Single words or groups of words can refer to entities, clauses can describe situations or events and texts can tell whole stories. One of the major difficulties in analyzing language is the fact that these different levels of meaning are not independent of each other. While the upward dependencies from words to sentences to texts seem obvious, also effects exist where the superordinate level influences the meaning of the underlying ones. For example, words can have multiple meanings and can only be distinguished by considering the sentence or text they occur in. Likewise, sentences can have multiple meanings and need to be interpreted within the discourse they occur in. Even whole texts can vary in meaning depending on the context they are published in. For example the meaning of the first verse of the Deutschlandlied depends on the temporal context. It changed heavily in what it represented ever since it became a symbol for Nazi Germany. Thus, the different levels of linguistic utterances are highly connected. Yet, most semantic analyses only work bottom-up.

Furthermore, there exist different linguistic fields for different levels and different types of analyses. Morphology analyzes the subelements of single words. Classical syntactic and semantic analyses work on sentence level. And corpus linguists concentrate on complete corpora. Many times, researcher from

one specific area stay within the boundaries of their field and treat situations where information from other levels might be needed as special cases instead of looking for a general solution.

Instead of separating linguistic levels and linguistic disciplines from each other, supersemantics promotes an integration of them. In this respect, the 'super' in the term stands for going beyond borders. This includes the artificial borders of the research fields as well as the linguistic levels of different utterances. The goal behind this is to always use the best available tool for the task at hand instead of feeling the need to stay within its limits. So, instead of performing a mere syntactic analysis based on possibly ambigue words, supersemantics would favor an analysis that integrates contextual lexical or semantic information that helps to overcome this ambiguity.

Summing up, one can state that supersemantics is grounded in three maxims that should be used as foundation when designing a linguistic analysis system: contextualization, integration of linguistic levels and integration of linguistic fields.

The rest of this chapter motivates the use of contextual information and bridging of linguistic levels. Furthermore, existing linguistic approaches that try to provide comprehensive analyses independent of linguistic levels are presented to describe the scientific background of supersemantic analyses. Finally, a brief overview of the levels of linguistic utterances and the relevant information that they contain is given as well as a road map on how one could integrate all of these aspects of a supersemantic analysis into one comprehensive all-in-one solution. The general picture that is given there is drawn in further detail in the following chapters. Each of these presents the details of a linguistic level and how it can be used to give valuable context information to tasks of lower levels. Additionally, at least one practical application that was implemented in the course of this work is introduced in each chapter. This way the practical relevance of the given information is shown. In the end, in chapters 9 and 10 first prototype implementations of text mining systems that could be extended to comprehensive supersemantic analyses are described.

2.1 Levels of Context

The English language is a very complex construct and in it meaning is transported on various levels. Furthermore, the different levels influence each other. Syntactic rules constrain the choice of words, semantic contexts clarify ambigue expressions. Different subfields of linguistics focus on specific levels or try to bridge them.

The most elemental unit in a written language is a letter. The smallest meaning-bearing one is the morpheme. Morphemes are either word roots or their prefixes or suffixes. For instance the word "friendly" consists of the morphemes "friend" and "ly". In some applications one is interested in mapping all mentions of words with the same root together. In such cases stemming or lemmatizing algorithms are used to reduce the word to this morpheme. A hierarchy of linguistically relevant levels can be seen in Figure 2.1. The figure is rather coarse-grained and could additionally include morphemes or chunks. However, for the approaches presented in this work the level of detail suffices.

The problems which Supersemantic methods attempt to solve bridge these different levels. This can be seen in Figure 2.2 where some typical linguistic problems are arranged according to the levels

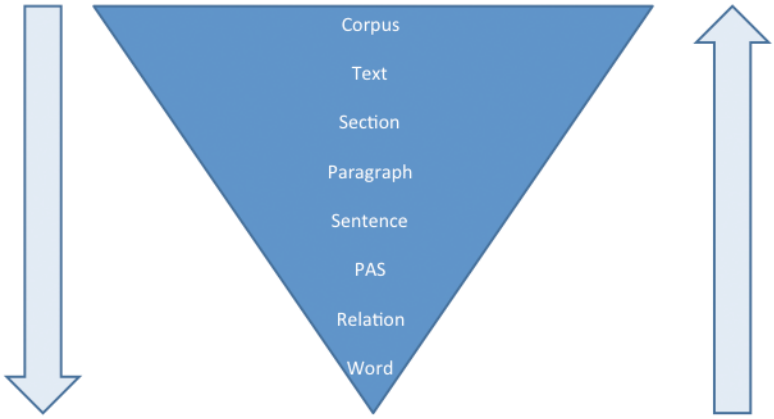


Figure 2.1: Different levels of linguistic utterances. The arrows indicate the interconnections between them. While classical approach usually only address the bottom-up direction, supersemantics tries to take both into account.

they affect. For example, for solving PP-attachment (the problem of deciding whether postposed prepositional phrases modify the verb or object of a sentence) one needs to either make logical deductions using external knowledge (e.g. in the sentence “Neprilysin is poorly expressed in dogs with cognitive dysfunction syndrome”, one can deduce that the prepositional phrase belongs to the object “dogs” using the external knowledge that “cognitive dysfunction syndrome” is something living creatures can have rather than being an instrument of gene expression) or make use of contextual knowledge from the same document (e.g. in the sentence “The man saw the boy with the binoculars.” one could check the rest of the text in order to find out whom the binoculars belong to).

Often there exist different approaches to solve a problem with respect to how much context is taken into account. It might be possible to disambiguate a word or to resolve an abbreviation by simply considering the sentence they occur in. But e.g. for abbreviations that are defined in the beginning of the text analyzing the whole text might be necessary. Of course one could as well steer a middle course by considering the paragraph or the section. The arrows shown in the illustration should only point to the most commonly applied ways to solve the problem and do not make a claim to completeness.

The different levels of language are also reflected in the linguistic disciplines that study them. Morphology deals with morphemes and affixes to determine the structural buildup of words. Lexicology concentrates on words and their meaning. Sentence analysis is the study of the structural rules by which sentences are formed. Text Linguistics deals with whole texts and the circumstances in which these texts are situated. Finally, Corpus Linguistics is concerned with the compilation and analysis of whole corpora of commonly used language. The last four of these different fields of linguistics were also included in Figure 2.2. It seems obvious that tools developed by researchers only focussing on one of these levels cannot avoid having problems with those problems that require bridging of levels.

Apart from these, however, there also exist linguistic fields that are not just focused on a single level or explicitedly focus on the effects of context. These fields are largely based on alternative theories to the Chomskyan Generative School. While Chomsky focussed on analyzing sentences, the alternatives

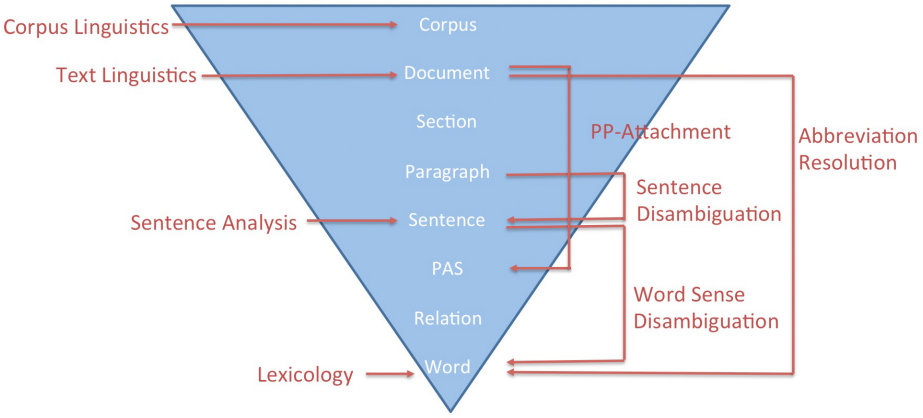


Figure 2.2: Different levels of linguistic utterances and how they need to be bridged in order to solve typical linguistic problems (on the right). Furthermore, some linguistic disciplines that focus largely exclusively on one of the levels are shown (on the left).

all believed that a comprehensive linguistic analysis should go beyond the sentence. This alternative disciplines that emerged in the twentieth century are (Alba Juez, 2009):

- Functionalism (functional grammars)
- Cognitive Linguistics
- Sociolinguistics
- Pragmatics
- Text Linguistics
- Discourse Analysis

The last three of these are the most relevant to Supersemantics and thus should be examined in the following. Discourse Analysis and Text Linguistics are related fields that focus on the analysis of complete discourses. They try to analyze intentions and focuses and other properties on complete texts. While the main focus of Discourse Analysis are spoken conversations, insights gained from it may also affect written text. Furthermore, Pragmatics deals with all aspects of meaning that are caused by the context and were not analyzed by the older field of Semantics. The following two sections will more comprehensively introduce these terms.

2.2 Pragmatics

The field of pragmatics is not easy to define. From its widest definition as "the study of understanding intentional human action" (Green, 1989) to its narrowest definition as "the study of [...] expressions whose reference is a function of the context of their utterance" (Green, 1989), there exist different

interpretations of the term. However, they seem to have in common that pragmatics deals with interpreting things that go beyond what is explicitely stated.

For example the sentence "The restaurant looks good." might be interpreted differently in different situations. It might be a proposal to enter the restaurant and a remark that one is hungry. If a restaurant critic states it on the other hand it might refer to the ambience and outer appearance of the establishment and if it is stated by an architect the architecture of the building might be referred to.

The wide definition of the term dates back to Charles Morris who first used the concept in modern times. In his study of semiotics (the science of signs), he distinguished syntax, semantics and pragmatics. While syntax described "the formal relation of signs to one another" and semantics described "the relation of signs to the objects to which the signs are applicable", pragmatics described "the relation of signs to interpreters" (Morris, 1938). In accordance with this distinction, he characterized pragmatics by stating that it "deals with [...] all the psychological, biological, and sociological phenomena which occur in the functioning of signs" (Morris, 1938).

While this wide interpretation of pragmatics lived on in fields like sociolinguistics or psycholinguistics (Levinson, 1983), the narrower interpretation of pragmatics is more common and is the one relevant for this thesis.

If one restricts oneself to the domain of language, pragmatics deals with believes, intentions and goals within the process of communication (Green, 1989). While semantics is the study of meaning, pragmatics is the study of language usage. In this connection, the interplay of language structure and the principles that govern how language is used are of special interest (Levinson, 1983).

In the view of pragmatics, when people communicate they do not merely exchange umambiguous information, but rather the recipient of a message successfully interprets the intent of the message. In doing so, a so-called linguistic act is performed (Green, 1989). This view of communication is based on Grice's theory of meaning (Grice, 1957, 1981).

In his framework, Grice sees the speaker as somebody who wants to communicate a certain communicative intent to a recipient. If the recipient understands the intent the communication was successful and the intent becomes mutual knowledge. Here, the intent of the speaker does not necessarily need to be formulated explicitely. For this reason, Grice distinguishes sentence meaning and speaker meaning. Pragmatics then is the study of explaining how sentence meaning can be translated into speaker meaning by considering the context. In the analysis of the meanings, one can make use of four maxims that ideally should guide the communication between speaker and recipient (Grice, 1957, 1981). These maxims are:

- maxim of quality (one only states accurate information)
- maxim of quantity (one is as informative as necessary, but not more)
- maxim of manner (one formulates comprehensibly to avoid misunderstandings)
- maxim of relevance (one talks about relevant issues with respect to the current topic)

While these maxims are not always consequently followed they can be used to formulate heuristics to resolve the speaker meaning (Blum-Kulka and Hamo, 2011). In the example given above, the meaning

of the sentence "The restaurant looks good." might be resolved by the maxim of relevance. If the sentence is embedded in a discussion of construction styles, this might be a good hint that the outer architecture of the building is referred to.

The analysis of linguistic acts - or more precisely speech acts - is based on the work of Austin and Searle. Austin was the teacher of Grice and introduced the term implicature for that what is meant but not said within an utterance. Searle's work was also based on that of Austin but in contrast to Grice focussed more on acts. Austin originally made a distinction of speech acts. Each act was characterized by being intentional and goal-directed by itself.

In such a framework, already the production of the sounds that constituted speech are acts on their own (phonetic act). The goal of this act is for the recipient to understand it and be able to use it to build the sentence it belongs to. A phatic act is the act of building an expression in accordance with the grammatical rules of the language. A rhetic act can be a reference to an object in the real world or a predication. A predication in turn is the process of assigning properties to a subject. With this process meaningful logical statements can be formed from text. Austin called these three kinds of acts locutionary. In contrast to this, an illocutionary act dinstinguishes the way in which a statement is used. This can be an act of stating, questioning, commanding, promising and so on (Austin, 1975).

Searle's speech acts were a bit more coarse-grained. He distinguishes between utterance acts, propositional acts and illocutionary acts. An utterance act subsumes everything that is necessary to utter words (producing the sounds of speech, forming a grammatically correct sentence, intonation, ...). Thus, he combines Austin's phonetic and phatic acts. The propositional act corresponds to the rhetic act in Austin's framework and Searle adopts the illocutionary act (Searle, 1969). Building on Austin's theory Searle formulated additional contextual conditions that need to be met in order for an act to be successful (Blum-Kulka and Hamo, 2011):

- Propositional content (properties of the semantic content of an utterance, e.g. requests reference s.th. in the future)
- Preparatory conditions (the necessary context information, e.g. a recipient must be able to perform the task requested from him)
- Sincerity conditions (the speaker's attitude, e.g. his wish that the recipient performs the requested act)
- Essential condition (the convention by which the statement is associated to the respective act, e.g. that the sentence "the restaurant looks good" can be interpreted as request to enter it)

While the theory of speech acts is more interested in a general theory of human communication, Supersemantics as described in this thesis focusses on the practical aspects. Thus, some preparatory conditions are of lower interest. At the current stage, one is interested in what is meant, not yet in whether this can also be realized. Also, the sincerity conditions can be assumed to be met in the context of scientific publications, even though the reliability of published results should be questioned in some cases. Such analyses fall into the domain of automatic reasoning, which is an important next step of future comprehensive Supersemantic analyses.

The findings of pragmatics build the foundation for a lot of Supersemantic methods. The propositional content and the essential condition are necessary to interpret texts. Also some preparatory conditions can be used to check whether an interpretation makes sense within that context. Likewise, the maxims of Grice are useful in situations where different interpretations of an utterance need to be distinguished with respect to the most likely one in the given situation. Additionally, the analysis of deixis is often also attributed to belong to the field of Pragmatics. As will be discussed in section 2.4, resolving deitic expressions is an important task for Supersemantics. Taking all of this into consideration one might think of Supersemantics as some kind of Computational or Applied Pragmatics.

2.3 Discourse Analysis and Text Linguistics

Discourse Analysis and Text Linguistics are related fields that root in different areas of research. While Text Linguistics is mostly studied by linguists, Discourse Analysis is applied in very different fields reaching from Anthropology over Sociology, Rhetoric and Literary Scholarship to Psycho- and Sociolinguistics (Alba Juez, 2009; van Dijk, 1977). While both disciplines usually analyze whole texts their focus differs. Text Linguistics aims at understanding the structure of coherent text. Here, so-called text grammars were developed to analyze texts analogously to sentences. Discourse Analysis, on the other hand, uses the analysis of texts as a mean rather an end. Its focus is the analysis of psychological or social factors that can be identified by analyzing the communication of people. For this reason, Discourse Analysis also often focuses on recorded speeches instead of written texts (Yatsko, 1998). Accordingly, Text Linguistics concentrates mainly on text-internal features like cohesion and coherence and is more formal whereas Discourse Analysis (sometimes also described as "the study of language in use" (Alba Juez, 2009)) concentrates more on external factors that help to put the text into context and explain its function within the general discourse. Furthermore, some fields started to apply Discourse Analysis also to non-verbal communication like gesture, dance, photography or clothing (Alba Juez, 2009; Titscher et al., 2000).

Since the differences between the two fields when dealing with text are more in nuances and focus than in substance, no distinction between the methods they use will be made here. Instead the term text analysis is used as an umbrella term refering to both fields. Among the most important phenomena studied in text analysis are coherence and cohesion. Both refer to different kinds of connections that need to be in place within a coherent text. Scientists started analyzing connections between sentences by looking at lexical and grammatical features. This syntactical analysis examines the cohesion of a text and is based on such phenomena as conjunction, ellipsis, anaphora and recurrence (Alba Juez, 2009; Yatsko, 1998).

However, it became clear that cohesion was not enough to describe all the connections between sentences in a coherent text. The sequence of sentences "It is summer. It is a table. It is difficult." is coherent with respect to its surface structure (the repetition of the subject) but does not make sense. On the other hand, the sequence of sentences "He wants to write a play for me. One act. One man. Decides to commit suicide." (from Bliss from Katherine Mansfield) lacks any form of syntactical connection but does occur in a real story and can be understood. Based on this the distinction between cohesion and coherence was established. While cohesion described the "relationship between text and syntax" (de Beaugrande and Dressler, 1981), coherence is concerned with the meaning of the text. This includes

semantic connections that are not explicitely realized by linguistic structures in the text (de Beaugrande and Dressler, 1981).

The possible semantic connections constrain the set the of acceptable sentences and would rule out the first of the two sequences. Van Dijk (van Dijk, 1977) dinstinguishes between two kinds of constraints: linear and global ones. While linear constraints arise from direct connections between sentences or parts of sentences like clauses, global constraints reach further. The global constraints are based on global structures which van Dijk calls macro-structures. Within these a sentence can contribute, like a word contributes to a sentence by taking a certain syntactic function. The difference is that the function of the sentence is a semantic one and that the concepts of macro-structures are less well understood than the syntax of a sentence. With these global structures van Dijk sees texts as more than merely sequences of sentences but moreover as a whole with its own structure.

A macro-structure can further be analyzed for its topic. Here, van Dijk based his framework on the more common analysis of the topic of a sentence. In discourse analysis, a sentence can be divided in topic and comment (also sometimes called focus). While the topic represents what is talked about, which often but not always coincides with the subject of a sentence, the comment represents what is said about the topic. Van Dijk extends this concepts to texts giving a whole text or part of a text a division into topic and comment. The topic of a text passage in turn constraints the comments within that text passage to match the topic. Moreover, topics and comments can be nested making it possible to create sub-topics. In line with this, he also points out the possibility to nest macro-structures (van Dijk, 1977).

Furthermore, a macro-structure can have a type that requires or allows certain sub-structures. In a novel e.g. the overall structure could be called a NARRATIVE which in turn subsumes the macro-structures SETTING (describing the setting of the story), COMPLICATION (describing the conflict), RESOLUTION, EVALUATION and MORAL. These sub-structures do not necessarily have to be at one stretch but can be discontinuous. For example the description of the conflict can be distributed over different parts of the novel. Different macro-structure categories can have different functions, e.g. the SETTING can have the function of introducing the characters, and these functions can effect the rest of the discourse like e.g. the SETTING can determine the language and location of the story (van Dijk, 1977).

Both Text Linguistics and Discourse Analysis are related to Supersemantics as presented in this thesis. The analysis of cohesion with phenomena like ellipses and anaphoras can increase the recall of a text mining system. Understanding conjunction, recurrence and coherence better will improve the precision by interpreting sentences correctly within their context. In the same way the constraints of macro-structures can help to better interpret ambigue statements by ruling out interpretations violating the constraints. Additionally, identifying topics of texts or parts of texts is a task that is included into the range of Supersemantic methods. Generally, the focus on language in use within Discourse Theory matches the requirements of biomedical text mining.

2.4 Why Supersemantics?

The reason why supersemantic methods are needed are manifold. Common errors of text mining systems occur because the context of relations or entities is ignored. The certainty of claims might be overemphasized because the speculative environment of the claim is overlooked. Or specialized questions could not be answered because the temporal or spatial restrictions of a claim are not captured. While pragmatics and text analysis cover some of these aspects, their scope is still too limited to focus on all of them. Furthermore, there are few approaches that practically realized computational solutions of these supersemantic tasks, let alone tried to integrate multiple or all of them in a single system. To give an overview of a selection of situations where supersemantics is needed a series of examples is presented in this section.

Disambiguation of terms

As mentioned before, words can have different meanings in different circumstances. This phenomenon is especially important in the biomedical domain where the frequency of ambigue terms is higher than in other domains. In order to understand which meaning is to be used the surrounding text can be analyzed. The meaning of the term can be given explicitly in the form of a definition or must be infered implicitly. Among linguists the following two sentences are a common example for ambiguity:

> "Time flies like an arrow. Fruit flies like a banana."

Here the word "flies" is used as a verb in the first and as a noun in the second sentence. Furthermore, "like" is first a preposition and then a verb. Gene names and symbols are often ambigue. Methods to distinguish the meaning could greatly increase the quality of relation extraction systems. The information needed to resolve ambigue terms properly is usually given within the surrounding sentence or text.

Abbreviation Resolution

Scientific terms are often long and composed of many words. In order to formulate more concisely as it is typically done in scientific publications and simply for the sake of convenience the use of abbreviations is very common. While sometimes standard abbreviations for certain terms like gene aliases exist, other abbreviations are introduced only within the scope of a publication. To resolve the used abbreviations properly one has to consider the complete text as well as additional sources where common abbreviations are stored.

The fact that this problem of abbreviation resolution is especially relevant in a context of scientific publications should be illustrated by Figure 2.3. In this example several abbreviations are nested resulting in a total of five levels of resolved abbreviations. While the term "V-SNARE" might be a special case with its many layers the use of simpler abbreviations is customary. Abbreviations could be resolved by the integration of external resource that already collected them. For example lexical resources like the Human Phenotype Ontology (Robinson et al., 2008) often have integrated lists of abbreviations and synonyms. In cases where the available resources do not cover the respective

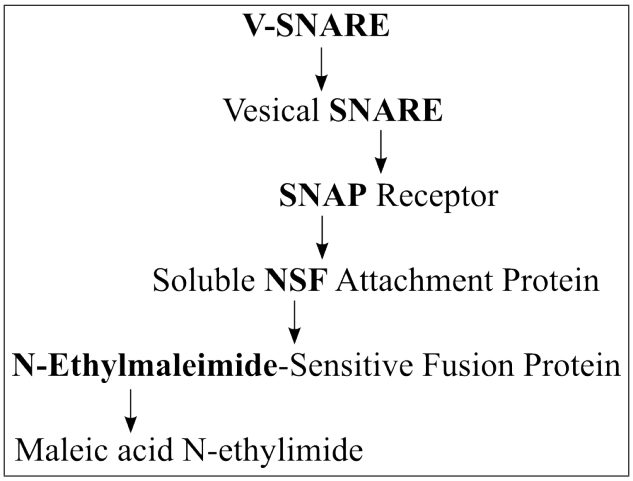


Figure 2.3: Different levels of abbreviations of the term V-SNARE. Figure based on (Morgan, 2005).

abbreviation, resolution methods can be applied. Commonly, an abbreviation is explicitly introduced within the scope of a text. Resolving this definition and using it as context information for the rest of the text analysis would be a supersemantic approach to this problem.

Deixis

Besides abbreviations there exist other elements of language that result in more concise phrasing. When writing a text the writer assumes some things to be known and introduces others within the text. Each of these can be referenced in a shortened way.

The most common practice here might be the use of pronouns. Apart from that, however, other pro-forms exist that reference other word types. Pro-adjectives refer to adjectives (e.g. the word “so” in “His is blue. So is hers.”), pro-adverbs to adverbs (e.g. “this way”), pro-verbs to verbs (e.g. “do”) and pro-sentences refer to whole sentences (e.g. “That’ in “That is true.”’). Furthermore, additional reference phenomena exist. The term anaphora when used in its wider sense subsumes all of them. The phenomenon that the interpretation of a word is dependent on contextual information is referred to by the term deixis.

Different expressions exist to reference things from different categories. The most common of these categories are the following:

- Persons: I, you, he, she, it, we, ...
- Objects: this, that, which, ...
- Places: here, there, this city, ...
- Time: now, today, tomorrow, earlier, later, ...
- Manner: hereby, thereby, ...

- Reason: herefore, therefore, ...

The resolution of what these expressions refer to is not a trivial task. The different kinds of references are unequally important to fact extraction. So focussing on the ones that contribute most seems to be the prefered strategy. In any case this problem is a very prototypical example of the urgent need of a contextualized text mining approach.

Disambiguation of Sentence Structures

Ambiguity of syntactic interpretations of sentences poses another problem. This came as a surprise for computer linguists when computers became powerful enough to implement the grammars they were creating. When presenting their system with the sentence "List sales of the products produced in 1973 with the products produced in 1972." Martin et al. (Martin et al., 1987) were presented with 455 different results. While the scope of this problem might be reduced by more sophisticated modern rule systems, the underlying problem remains. Syntax on its own cannot avoid ambiguity.

More intuitive examples of these kinds of problems are given by common linguistic example sentences:

> "Police help dog bite victim."
>
> "The man saw the boy with the binoculars."

From a syntactic point of view it seems equally likely that the police helped the dog or the victim. Furthermore, it is not clear if the man used the binoculars to see the boy or if the boy was the one with the binoculars. To overcome this form of ambiguity additional context information is needed. This problem mainly boils down to choosing the most sensible interpretation that does not violate any linguistic or logical constraints. In order to measure this meaningfulness a reasoning mechanism based on the information extracted from the text and possibly additional a priori knowledge would be necessary.

Negation and Speculation

Relation extraction focuses on the verb, subject and objects of a sentence. This approach on its own is ignoring negations and speculations. While it is obvious that negations are essential to a properly working text mining system also the detection of speculations holds great value.

Within the discussion sections of papers scientists tend to speculate about further implications of they results and possible future connections that still need to be verified. If these kind of relations are extracted by a relation extraction system the results become less reliable. Thus, identifying both negations and speculations may improve every text mining system.

Additional Contextualization

Even if a relation is extracted properly the context of its validity has to be considered. A biological event might only occur in certain species, tissues or cells. Results obtained from experiments with mice

might not always be transferable to humans. Furthermore, the time and manner might be important. For example, the method by which the results were obtained might have implications on the reliability of the result.

Negations, speculations and other contextualizations are important to get a more fine-grained picture of what was said in a text. Depending on the application it can be crucial to add these kind of information in order to get meaningful results.

Argumentation Analysis

A more advanced application for supersemantics might be the analysis of the argumentation structure of a text. Humans are prone to logical fallacies. Causation is often implied by correlation, statements are considered valid because of the reputation of the person stating them, cause and effect are mixed and probabilities are interpreted incorrectly. There exists a long list with dozens of typical fallacies humans tend to fall for. An automated system that is capable of detecting such errors could be of benefit for identifying unsufficiently supported claims. The challenge here is that arguments usually span over several sentences. Thus, a supersemantic application incorporating a multi-sentence analysis would be required.

All of these problems fall within the range of supersemantic applications. In each of the cases information from multiple linguistic levels is integrated or used to complement the analysis results. It is the goal of this thesis to work towards solving them.

2.5 A Supersemantic Analysis

In order to solve these problems a more sophisticated analysis than the typical text mining pipeline shown in Figure 1.7 is needed. The existence of both bottom-up and top-down dependencies excludes the possibility to comprehensively analyze text in a mere sequential manner. Thus, a supersemantic analysis replaces the typical processing pipeline by an architecture that allows connections in both directions. How such an architecture might look like is shown in Figure 2.4.

Here, the pipeline is substituted by a network setup. As one can see, different levels of utterances (words, sentences, sections, texts and corpora) are arranged in different columns. On each of these utterances different analysis steps (arranged in the corresponding columns) are based. The network structure, however, allows connections from higher level procedures to lower level ones. The protypical architecture given here does not make any specifications on how to realize such backwards connections. Possible realizations, however, include correction procedures that trigger reprocessings from the point-of-change onwards, multi-objective optimization techniques that try to maximize the utility of the different constraints, that arise at the different stages of the linguistic analysis, all at once, and multi-task machine learning algorithms that optimize a variety of objective functions simultaneously.

While Sentence Splitting and Tokenization are left out of the scheme for readibility, the other steps in the prototypical text mining pipeline can be found in Figure 2.4. The named entitiy recognition is placed in front, since it creates the foundation for the sentence analysis. Furthermore, it is extended to

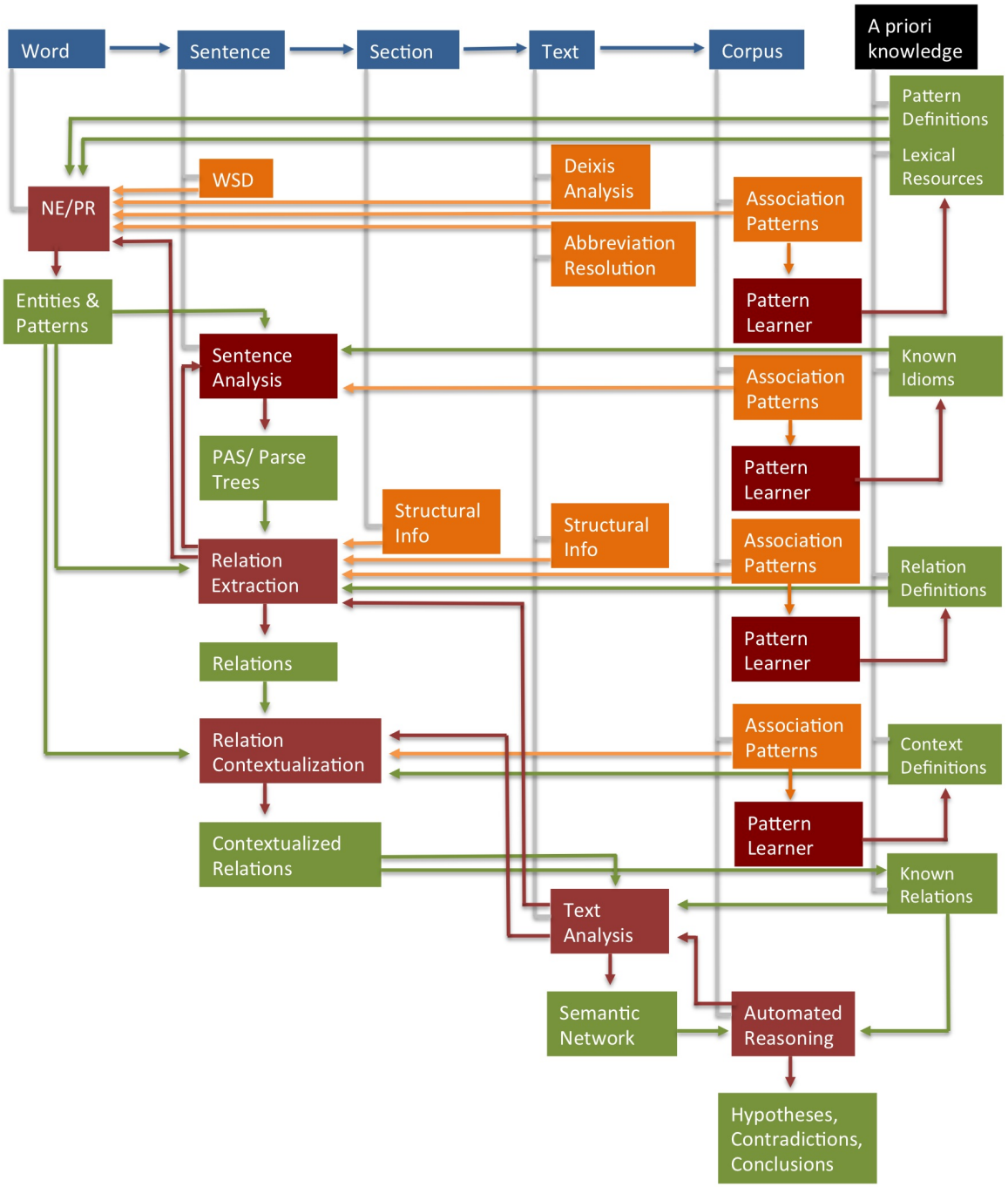


Figure 2.4: A prototypical overview of a comprehensive supersemantic analysis network.

include a pattern recognition step. Thus, it is referred to by NE/PR , named entity/ pattern recognition. These patterns include utterances that can be detected without considering contextual information, like time designations, or domain specific patterns like citations in scientific publications. Placing these steps in front allows an early integration of semantic information into the otherwise mere syntactical analysis, thus helping to bridge the distinction between these two linguistic fields.

Also for readibility, POS tagging and sentence analysis are subsumed to a single node called sentence analysis. Finally, the last step, event extraction, is included in the analysis network. It is combined with the prior relation recognition in the node called relation extraction. In addition to these, the supersemantic analysis includes further steps. The extracted relations are first contextualized and then combined in a text analysis module. Finally, the semantic networks of contextualized relations from many texts are combined with knowledge from external resources in an automated reasoning module. This module then produces hypotheses, reveals contradictions and is able to draw conclusions.

Apart from these additional processing steps, a second difference to the typical text mining pipeline is the inclusion of several tools that use higher level analysis to improve lower level ones. On a sentence level, a word sense disambiguation (WSD) module supports the NE/PR module. On section and text level, information about the structure of the document supports the relation extraction. Furthermore, on a text level, the resolution of abbreviations and deitic expressions additionally facilitates the NE/PR. Finally, on corpus level, different association patterns are collected. Such association patterns are frequently occurring patterns that can be used to improve linguistic analysis on several levels. If e.g. an association pattern of a typical formulation “as mentioned by x” where x is always a person is found, then this can be used in the NE/PR stage to distinguish different meanings of words. Thus, if the word “Pidd” occurs in an expression “as mentioned by Pidd” the NE/PR module can infer that here a person called Pidd is meant and not the p-53-induced death domain protein PIDD. Such patterns can then in turn be used to extend one’s external resources. In the previous example, the term “Pidd” could be included in the lexical resource used of NER as a person if it would not have been known before. Such a procedure is called Bootstrapping. In Figure 2.4, the more general name pattern learner is used to indicate that various methods are possible. Analogously, association patterns could be used to improve relation extraction and relation contextualization (association patterns will be presented in more detail in section 7.1).

The work presented in this thesis is directed towards the realization of such a supersemantic analysis network. While the realization of a complete system is beyond the scope of a single dissertation, different modules were implemented in the course of this work. Chapter 3 describes a WSD system. Chapter 4 gives additional information about various ways of relation contextualizations and describes a relation contextualization module that complements relations with information about whether they are negated. Chapter 5 explores the ways in which structured information from semi-structured documents can be included in the text mining process and describes an analysis in which such information was used to analyze semi-structed articles about rare diseases. Chapter 6 provides further information about contextualizations based on whole texts, like abbreviation resolution systems. Additionally, it introduces an anaphora resolution module (a type of deixis analysis). Chapter 7 provides an overview of corpus-based contextualizations and describes a visualization tool that is based on word association patterns (a so-called word space model). Next, chapter 8 talks about the integration of external knowledge with text mining results and describes a tool using such an integration in order to facilitate the interpretations of gene set enrichment analyses. Finally, chapters 9 and 10 describe the first prototypes

developed in the course of this thesis that could serve as the foundation for a supersemantic analysis network like the one shown in Figure 2.4.

2.6 Related Work

An overview over related related approaches that try to provide a comprehensive, context-aware analysis of texts is given in table 2.1

Table 2.1: Supersemantics: Related work

Authors	Year	Approach	Domain
Gerner et al. (2012)	2012	Contextualized event extraction	Biomedical
Mei (2009)	2009	Contextual text mining	General
Csaba (2013)	2013	Contextual named entity recognition	Biomedical
Pecheux (1995)	1995	Automatic discourse analysis	General
Evi (2014)	2014	Pragmatic question answering	General
Graesser et al. (2004)	2004	Text linguistics for measuring text coherence	General
Brown et al. (2008)	2008	Text linguistics for measuring idea density	General
Ferrucci (2011)	2011	Pragmatic question answering	General
Van Landeghem et al. (2013)	2013	Contextualized event extraction	Biomedical
Tamames and de Lorenzo (2010)	2010	Contextualized measurement extraction	Biomedical

3

Sentence Contextualization

Supersemantics is an umbrella term that encompasses many methods that look at the context of utterances to better understand their meaning. To exemplify the importance of context and at the same time to tackle one of the big problems in text mining this chapter introduces a word sense disambiguation method that ignores the word that should be disambiguated completely and instead infers it's meaning solely from the context.

The importance of disambiguation algorithms in the biomedical domain was already pointed to in the previous chapters. Especially, genes and gene aliases are often ambiguous and context information is necessary to distinguish them from common english words or abbreviations of compounds, diseases or other biological terms. This is a core application field for Supersemantic methods, since meaning is deduced from context. The algorithm realized in this thesis formulates the problem as a classification problem and utilizes a spam filter to distinguish between meanings. It was designed together with Anita Winkler who implemented it in the course of her Master's Thesis (Winkler, 2011). I contributed to this project by modifying the problem formulation into a two-class-problem, extending the approach to contain the different pre-processings, and supervising the development process.

3.1 Ambiguity and the Need for Disambiguation

Words, sentences and whole texts can vary in meaning when uttered in different tonality or context. Examples of this are ambigue words, ironic or metaphoric statements, puns, vague statements and many others. Linguistics distinguishes different kinds of meaning variations. If the definition of an utterance is general enough to apply it to many different things, then one speaks of vagueness. For example the term clock can refer to all kinds of clocks from digital ones to cuckoo clocks. If two meanings just happen to have the same form, one speaks of ambiguity. This is the case for sentences

or words with multiple meanings. Finally, polysemy is a mixture of the two. If an utterance has two different meanings which, however, are related to one another, one calls this polysemy. For example the word 'foot' can refer to the body part or the scale unit. Both meanings, however, have the same root in the Old English word 'fot'.

The different distinctions between these meaning variations are not clear cut but one can rather consider them on a continuous scale with ambiguity and vagueness at its extremes (Deane, 1988; Murphy, 2010). All kinds of meaning variations are addressed in this thesis, starting with ambiguity in this chapter.

Ambiguity is omnipresent. It occurs on all levels of linguistics. Beside the already mentioned semantic, syntactic and lexical ambiguity, morphemes (e.g. "'s" can indicate a genetive or the short form of is) and phonemes (the sounds used to build words) can be ambigue. Ambiguity, however, must not be considered an imperfection or shortcoming of a language. Instead, linguistic studies show that the use of ambigue terms can improve effective communication (Piantadosi et al., 2012).

The occasional use of ambigue utterances is a tradeoff between clarity and ease. In a communication, a speaker wants to minimize his effort in the production of language (he wants to use less, shorter and, if the communication is verbal, easier to pronounce words) to tranfer a certain message to a recipient. He can achieve this by reusing expressions that are easy to utter for different meanings as long as the recipient still understands what was meant. The number of easy to communicate utterances, however, is limited. This is why they are most often ambiguous. An example where ease is traded for clarity is the use of pronouns which are easier to express for a speaker but require deduction from the recipient. An example where clarity is traded for ease on the other hand is the NATO alphabet where letters are replaced by whole words to ensure greater clarity (Piantadosi et al., 2012).

In order to maximize both aspects one leaves out information that can be inferred from the context. This way less redundant information is given while at the same time the clarity is largely unaffected. Generally, there is a tendency towards ambiguity since for humans articulation is considered expensive while inference is considered cheap (Piantadosi et al., 2012). For computers on the other hand production is easy while inference is complicated. This asymmetry explains why disambiguation is a major problem for text mining.

Armstrong (Armstrong, 2010) points out that the main problem in the biomedical domain are ambiguous gene aliases. Beside the official name, genes usually have a set of aliases that serves as abbreviation of the full name. The use of this aliases is very common. Schuemie et al. (Schuemie et al., 2004) report 70% in abstracts and 82% in full text. At the same time these aliases are way more ambigue than full names as suggested by the tradeoff between ease and clarity. Chen et al. (Chen et al., 2005) report for mouse genes an ambiguity rate of 14% for full names and 85% for aliases.

While one might intuitively expect that the use of aliases is decreasing over time in favor of the use of a controlled vocabulary, Armstrong found that the opposite is the case. The use of multiple aliases is increasing. Armstrong tries to explain this by arguing that the different aliases may serve different information needs. If a certain alias is established in a certain domain or for a certain organism, it is e.g. easier to find information about the entity within this specific context, if a special name exists for this. This finding suggests that the use of aliases and, as a consequence of this, the ambiguity of genes will remain an issue in the future.

As already mentioned especially lexical ambiguity poses a major problem for biomedical text mining. Falsely resolved ambigue terms lead to the extraction of incorrect biological events, which if used in a systems biology model leads to incorrect edges between biological entities. Furthermore, if the same word can have multiple part-of-speech tags (e.g. 'fly' can be a verb or a noun), incorrect disambiguation can lead to an incorrect syntactic analysis of a sentence. Thus, the error gets propagated to further levels of the linguistic analysis.

3.2 Context Matters - The Case of Word Sense Disambiguation

Word sense disambiguation (WSD) is a classification task. Depending on whether a limited set or all words are disambiguated the task is called 'Targeted WSD' or 'All-words WSD'. A variety of approaches has been proposed to tackle the problem: different supervised learning algorithms (decision lists, decision trees, naive bayes classifiers, neural networks, instance-based learning, support vector machines, ensemble approaches and others), unsupervised (usually clustering word meanings without a predefined set of meanings) and knowledge-based approaches (making use of external resources like dictionaries, thesauri or ontologies) were employed.

The main quality measures for WSD systems are precision, recall and F-measure. Precision is the fraction of correctly labeled terms over all labeled terms. Recall is the fraction of correctly labeled terms over all terms that should be labeled. And the F-measure is a combination of the two defined as: $2 * \frac{Recall * Precision}{Recall + Precision}$. Furthermore, Accurracy (the fraction of correctly labeled terms) is often taken into consideration. In addtion to these quality measures the applicability to large corpora is important. For this reason, a good disambiguation system should be scalable to large amounts of data and be able to process them quickly. Additionally, generic methods are preferred over specialized ones, since they can be reused for similar problems. These performance measures, of course, do not only apply to word sense disambiguation algorithms, but serve as quality assessments throughout this thesis for all implemented modules and prototypes.

The approach presented in this work is a Targeted WSD system realized by a generic supervised learning algorithm that has been shown to outperform comparable approaches in the field of spam filtering. It was designed with focus on high quality classification and the good applicability to large scale problems.

3.3 Approach

In the approach presented here, the words that should be disambiguated are ignored in the classification process. The decision about the word sense is solely based upon the context of the word. In accordance with this the disambiguation turns into a text classification task of the context. A common field of text classification is spam filtering. Here, the CRM114 (Yerazunis, 2004) has proven itself as the most successful framework. The underlying idea of the presented WSD system is to take advantage of the reformulation of the problem as text classification problem and to apply the CRM114 to the task.

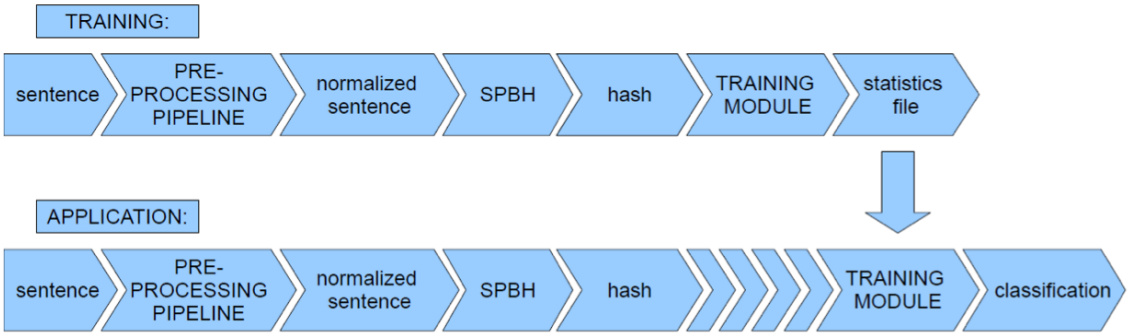


Figure 3.1: Workflow of the biomedical WSD system. Picture taken from (Winkler, 2011).

Figure 3.1 shows how the preprocessing steps and the CRM114 are integrated into a processing pipeline to train and test the classifier. The particular steps in the classification process are described in the following.

Data Set Generation

In order to apply and evaluate a supervised learning algorithm, train and test data sets are required. To maximize the use of the trained classifier and the relevancy of the evaluation the categories with the most ambigue biological terms were determined. To this end, eight broad categories (Gene, Compound, Disease, Pathway, Organism, Environmental Factor, Gene Ontology Categories, Others) were defined and filled with terms from 17 different databases. The four categories with the highest level of ambiguity (Gene, Compound, Disease, Others) were chosen and used for further analysis.

The train and test data sets were designed to include common, highly ambigue terms. To achieve this, the number of categories a term can belong to (as indicator of ambiguity) as well as the number of databases it occurs in and the frequency with which it occurs in MEDLINE (as indicator of commonness) were considered. On the basis of this list the data sets were created by searching sentences containing one of the top-ranked terms and annotating it to one of the four chosen categories.

This process resulted in two data sets. Data set Manual I contains of 168 sentences from 14 different terms from the four categories. It was extended to data set Manual II to be uniformly distributed between all classes. Manual II consists of 240 sentences. Furthermore, the test set of the second data set (T2) is based on different terms than the corresponding training set, to test the generalizability of the approach.

Preprocessing

Different preprocessing steps were applied in order to increase the generalization abilities of the features learnt by the classifier. The text was first tokenized using a regular expression. Additionally, case folding and stemming were applied. Case folding turns all words into lower case. This was done to recognize words when they were capitalized (e.g. at the beginning of a sentence or in headlines).

Stemming reduces the word to its stem by removing the ending. The ending of a word usually encodes grammatical but no semantic information (e.g. 'activates' and 'activating' both have the stem 'activat', which contains the semantic information, while the ending only encodes the tempus). By focussing on the word stem the same features were created for semantically equal words in different grammatical forms. This was done to improve the generalization properties of the classifier.

Furthermore, in term labeling an unusual preprocessing step was included. The classifier constructs its features from word combinations of the sentence. For a sentence like 'The patient suffers from diabetes' e.g. it will build features from combinations like 'suffers from diabetes' or 'The patient _ _ diabetes'. Such features however would not match for combinations like 'suffers from Alzheimer' or 'The patient _ _ Raynaud syndrome'. By replacing the disease with a dummy term, however, these combinations would turn into 'suffers from <dummy term>' and 'The patient _ _ <dummy term>'. In this case the features derived from the combinations are equal. Term labeling was used to avoid the classifier from simply memorizing words and instead to force it to create more general features.

CRM114

The CRM114 is a powerful classification framework. Instead of consisting of a single classification algorithm, it offers a multitude of classifiers (Markovian (MV) , Orthogonal Sparse Bigram (OSB) , OSB Unigram (equivalent to Bayesian classifier), OSBF (OSB + frequency features), Winnow, Correlate and Hyperspace) and training methods (train everything (TET) , train only errors (TOE) , single sided thick threshold training (SSTTT) and train until no errors (TUNE)), both of which can be flexibly used to combine them to powerful classification algorithms (Yearzunis, 2006). For the biomedical disambiguation system a workflow consisting of a Markovian or OSB classifier using the TET method was implemented and applied to 2-, 3- or 4-class classification problems.

Both Markovian and OSB classifiers are Bayesian filtering techniques based on a method called Sparse Binary Polynomial Hashing (SBPH) in combination with the Bayesian Chain Rule (BCR) . SBPH is used for feature generation from text. While classical Bayesian text classifiers consider each word as a feature, SBPH uses a sliding window approach where combinations of words from such a window are the basis for feature creation. In general SBPH works with arbitrary window sizes. The Markovian and OSB classifiers of the CRM114, however, use a window size of five. As seen in Figure 3.2 for a window size of five 16 different combinations of words are created. Each of these combinations consists of one to five words. 32-bit hashes are then created from these combinations and subsequently used as features for the classification. In this step the OSB classifier differs from the Markovian by only considering combinations from the sliding window that consist of two words. This way the method takes less time to process, requires less memory and has also often been found to perform better than the Markovian classification (Yearzunis, 2006).

Depending on the number of occurrences of a feature within a certain class, probabilities for each class given the feature can be calculated:

$$P(C|F) = \frac{P(F|C) * P(C)}{(P(C|F) * P(C)) + (P(\overline{C}|F) * P(\overline{C}))} \tag{3.1}$$

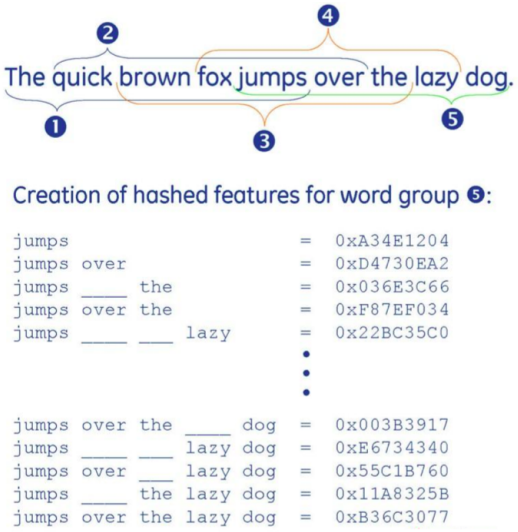


Figure 3.2: Example feature selection process using SBPH. Picture taken from manage-this.com [5].

Here, the probability that a given text that shows feature F belongs to class C is given by the a priori probability of C ($P(C)$) multiplied by the conditional probability $P(F|C)$, which denotes the probability that the given feature occurs when a text from class C is given and which can be inferred simply by couting the feature occurrences of the classes. The term is normalized by $P(C|F) * P(C)) + (P(\overline{C}|F) * P(\overline{C})$ to account for the feature occurrences in other classes($\overline{C}$). In the CRM114 implementation the conditional probabilities $P(F|C)$ and $P(F|\overline{C})$ are bounded to avoid probabilities of 0 and 1. Using the Bayesian Chain Rule with the Bayesian assumption that the features are independent, the probabilities of the single features can be combined to a conditional probability of a class given the whole set of features of a text by multiplying them .

Even though the Bayesian independence assumption is clearly violated in this case, the classification process still performs well. However, the probabilities suffer from the dependence between the features, which results in overconfident classifications. To circumvent such misinterpretations the CRM114 offers so-called pR values that should give a better assessment of the classification confidence. These confidence values are calculated by the difference of the logarithms of the probabilities that the given text is within (P) or outside ($\overline{P}$) the given class:

$$pR = log_{10}(P) - log_{10}(\overline{P}) \tag{3.2}$$

From the different training methods offered by the CRM114, TET turned out to give the best results (see Appendix A). This method is the standard training method that takes all training examples equally into consideration.

[5] http://manage-this.com/taglocity-autotag-engine-crm114, Accessed: April 24th 2014

3.4 Results

The WSD system was evaluated on different problem formulations, configurations and data sets to give a comprehensive performance overview. Furthermore, the effect of the different preprocessing steps was analyzed. A complete overview of the evaluations is given in Appendix A. The most important findings are described in this section.

First, the effect of classifying sentences instead of complete abstracts was evaluated. It turned out that using sentences as basis for the classification process increased the accurracy by 2%-8% depending on the problem formulation and data set. Replacing the word by a dummy term increased the accurracy by 2% on data sets that consisted of terms not used in the training set and decreased it by 3%-9% in those that contained the same terms as in the training set. This indicates that the classifier tends to learn certain terms by heart without the term replacement. When replacing the term, however, a better generalizability was achieved. Case folding improved the accurracy by another 2%-4% on the test sets with different terms and stemming by another 3%-4%. In total the preprocessing steps were responsible for a performance increase of 11% on the four class problem and 17% on a two class problem (gene vs. disease) given the data sets with different terms. This showed the usefulness and great importance of the applied preprocessing steps.

Using its optimal configuration the system was evaluated on T2. The configuration consisted of all the described preprocessing steps, the OSB learning algorithm, and the TET train mode. An accurracy of 88% was reached. Since T2 is a balanced data set, also the performance for imbalanced data was tested. To achieve this, half of the gene samples of T2 were taken out. Such a 25% reduction of the data set size naturally leads to decreased accurracy values. However, still an accurracy of 76% was accomplished.

In order to compare the performance of the WSD system with other state of the art systems, it was further evaluated on the BioCreAtIvE task 1A competition data set. BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) is a community-wide challenge that evaluates biomedical text mining systems. It was initiated in 2003 to establish common standards and evaluation criteria in the thriving field of biomedical text mining. The first challenge consisted of three tasks:

- **Task 1A:** A named entitiy recognition task. Gene or protein names had to be detected in unstructured text.
- **Task 1B:** A gene name identification and normalization task. The gene names found in an abstracted had to be mapped to unique identifiers from different sources.
- **Task 2:** An event extraction task. Gene Ontology terms together with the corresponding found proteins had to be returned from full-text articles.

Since word sense disambiguation is a crucial part of named entity recognition, the first of the three tasks was chosen to compare the system with other contestants. The BioCreAtIvE corpus used for task 1A consists of a training set of 15.000 sentences and a testing set of 5.000 sentences. It was derived from texts in MEDLINE and annotated manually. Both training and test data sets are nearly evenly split between sentence containing a protein name mention and sentences not containing one. To apply the WSD system to a NER task it was extended with a gene vocabularity. For this purpose the Entrez Gene database was used. The evaluation was performed on the BioCreAtIvE corpus reduced to gene

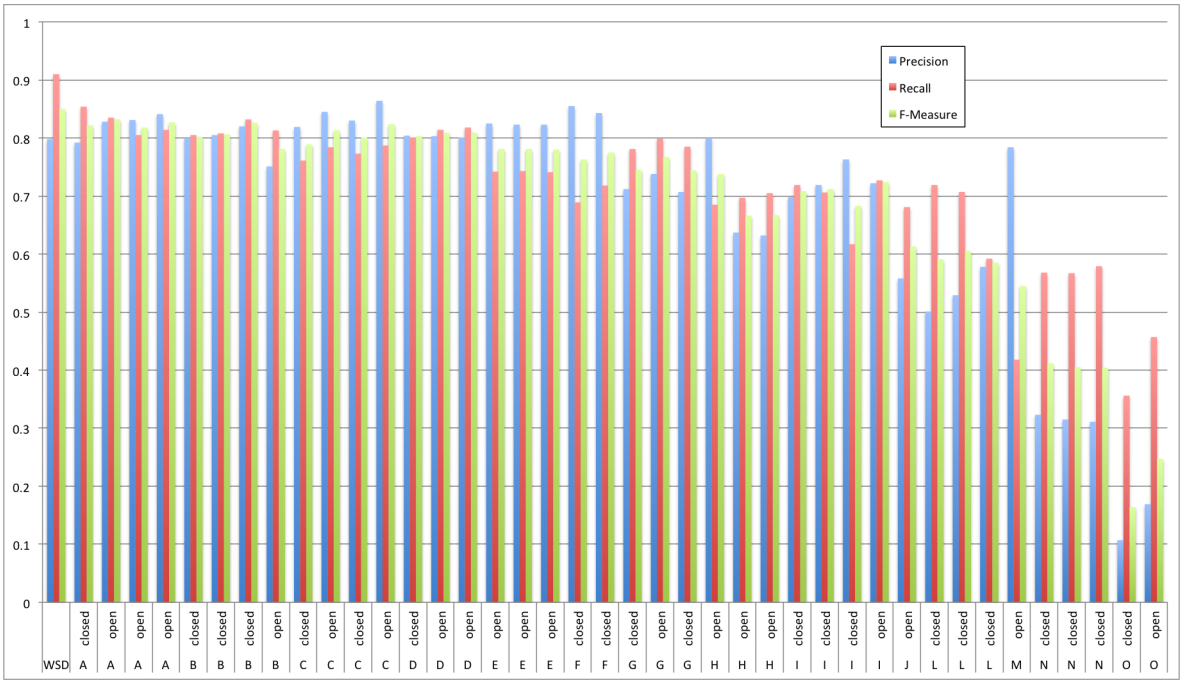


Figure 3.3: Comparison of WSD-based NER with contastants of BioCreAtIvE task 1A. The results for all 15 participants of the BioCreAtIvE task (A-O) for the open and closed version of the task are shown.

mentions that actually occured within Entrez Gene. A two class classification algorithm was trained on the classes gene and others. The OSB and Markovian classifiers were both used in TET mode. For comparison with the results obtained in the BioCreAtIvE challenge precision, recall and F-measure values were computed. The results are shown in Figure 3.3.

The WSD-based NER system reaches a recall of 0.91, a precision of 0.798 and a f-measure of 0.85. The numbers correspond to the Markovian classifier, which performed slightly better than the OSB classifier that reached a f-measure of 0.844. As can be seen, the WSD system outperforms all other closed and open systems of the challenge with respect to recall and f-measure. The term open system referes to algorithms that make use of external resources like word lists, while closed systems do not. The displayed letters refer to the different teams, which were allowed to submit mulitple solutions.

Since the obtained results were very promising, the WSD system was integrated into the Excerbt text mining system. The approximately 130 million sentences of Excerbt were analyzed with OSB and Markovian classifiers and the results were stored in Excerbt's database. The system operated at a speed of 5000 sentences per second, which showed its usefulness for application to large scale problems.

3.5 Conclusion

The adaptation of the software to the biomedical domain was successful. It could be shown that the necessary semantic information could be deduced solely from the context. The system could

outperform comparable approaches and delivered results that could be used to improve existing text mining systems. One of the most important steps in achieving these promising results was the use of the different preprocessing steps, which improved the performance by over 10%. In a comprehensive supersemantic analysis, the disambiguation of additional word classes, sentence structures and relations should be included to account for all levels of ambiguity in natural language.

3.6 Related Work

Table 3.1 gives a brief overview of some of the most important related approaches to word sense disambiguation.

Table 3.1: Word Sense Disambiguation: Related work

Authors	Year	Approach	Domain
Al-Mubaid and Gungu (2012)	2012	Suppor Vector Machine	Biomedical
Alexopoulou et al. (2009)	2009	Ontology-, Metadata-based	Biomedical
Hatzivassiloglou et al. (2001)	2001	Supervised Learners	Biomedical
Pedersen et al. (2003)	2003	Network Similarity Measures	General
Pedersen (2006)	2006	Unsupervised corpus-based	General
Murata et al. (2001)	2001	Bayes, SVM	General
Escudero et al. (2000)	2000	Naive Bayes, Exemplar-based	General
Garla and Brandt (2012)	2012	Semantic Similarity Measure	Biomedical
Preiss and Stevenson (2013)	2013	Semantic Similarity Measure	Biomedical
Chen et al. (2013)	2013	SVM-based active learning	Biomedical

4

PAS Contextualization

A relation between entities is commonly defined by the interacting entities and the relation type. Such a definition, however, does not account for all relevant nuances of meaning. A statement described by a relation might only hold under certain temporal or spatial constraints. It might be speculative or a fact. To capture this additional information, the contextualization of relations is important. Many contextualizations are given within predicate-argument-structures. The probably most fundamental of these is the distinction whether or not a relation is negated. In this chapter, a brief overview over possible uses of information from a PAS is given. Furthermore, the integration of a negation contextualization into Excerbt and its application to build a resource of negated protein-protein interactions is described.

4.1 Predicate-Argument-Structures as Context

A good first approach of trying to give an overview over all the information contained in a predicate argument structure is looking at the different roles that can be a part of one. The Propbank definition of roles lists 24 roles (see Table 1.1). This contains the relation type and entities directly involved in the relation like the agent (who does something), theme or patient (whom something is done to), instrument (with which something is done) and benefactive (who receives something). These correspond to the common definition of a relation (though they can include more than two players). However, there exist additional roles that describe the certainty of the relation (negation, speculation), and add further contextual information about the place, time, manner, extent, cause, purpose and further more detailed information. All these roles can be analyzed and used to complement the relation to give a more precise and more informative picture. The integration and application of the negation information is described in this chapter.

A PAS can further contain implicit information. In such a case, it can e.g. be used to resolve certain types of lexical ambiguity. Exemplarily, the term ”big brain” can refer to an anatomical description. When it occurs in a PAS as an ARG1 of the predicate ”express” one can, however, deduce that the gene with the same name is meant. Such information can be based on implicit associations of predicate-argument pairs e.g. by considering co-occurrence statistics or training machine learning models. Alternatively, patterns of certain biological events can be formulated explicitely. Such patterns are used in the integrated text mining system presented in chapter 10.1 and based on these the event feature of the anaphora resolution algorithm presented in chapter 6 was designed.

4.2 Negative results

Protein-protein interactions (PPIs) are omni-present in the human body. They are involved in most biological processes such as gene expression, cell growth, signal transduction, apoptosis and many more. Because of their central role they are a major focus of research. Different resources exist that try to collect all known interaction. The IntAct database (Kerrien et al., 2012) currently contains nearly 50 000 interactions. Stumpf et al., however, estimated the total amount of interactions in humans to be around 650 000 (Stumpf et al., 2008). Taken that the estimate is appropriate and that IntAct appropriately covers the current knowledge about PPIs, this means less than 8% of the existent PPIs are currently known. Correspondingly, research in finding more interactions is ongoing. One approach to this is the use of prediction algorithms that try to hypothesize PPIs based on structural, phylogenetic and other features. Commonly, these algorithms are supervised machine learning classifiers like random forests or support vector machines.

In order to use supervised learners, training data is required. Thus, protein pairs known to interact and those known not to interact are needed. While resources for known PPIs existed for longer until the creation of the Negatome (Smialowski et al., 2010) there was a lack of resources for non-interactions. Consequently, often simply random pairs of proteins, for which it was not known whether they interacted or not, were used as negative samples. The Negatome database could offer an alternative for this imprecise approach and was already used for this purpose (Valente et al., 2013). Futhermore, it was used in the classification of structural features of interaction interfaces (Planas-Iglesias et al., 2013), benchmarking high-throughput experiments (Hosur et al., 2012; Royer et al., 2012) and the conduct of network-based gene function inference (Erten et al., 2011). Because of its different uses, it was also integrated into IntAct.

The Negatome database was created by analysis of the three-dimensional structure of protein complexes and by manual annotation, the latter of which is very time-consuming. Thus, to accelerate the second release of the database, text-mining was employed to facilitate the annotation process. For this purpose, Excerbt was extended to be able to find pairs of proteins that have been described in literature to be non-interacting. This work was done in collaboration with Dmttrij Frishman, Florian Goebels, Pawl Smialowski (IntAct filtering, merging of data sets, structural analysis), Goar Frishman, Andreas Ruepp (manual annotation), and Benedikt Wachinger (PAS extraction). I implemented the filtering of the text mining results, designed and implemented the confidence score and the annotation tool, evaluated the system at the different stages, analysed the ability of Excerbt to reproduce Negatome 1.0, and

performed the error analysis. The results were also published in the database issue of Nucleic Acids Research (Blohm et al., 2014).

4.3 Extraction of Non-interacting Protein Pairs

The relations extracted by Excerbt were extended by a possible negation modifier by extracting the ARGM-NEG role returned from Senna for the corresponding predicate-argument-structure. Thus, in this case the PAS was considered as context for the biological event and used to complement the information of the event. These extended relations were then used to detect non-interactions between proteins by filtering out any PAS that did not contain an ARGM-NEG role or did not contain proteins in the ARG0 and ARG1 roles.

This resulted in 58733 potential non-interactions. In order to get an idea of the quality of the found candidates a sample of 20 sentences was inspected manually. Within this sample set only 20% were actual non-interactions. Analyzing the misclassified non-interactions suggested that the low accurracy might be due to verbs not describing an interaction.

Thus, to increase the precision of the filter, in a second iteration the verb of the PAS was additionally constrained. A narrower set of allowed verbs describing interactions or bindings (to interact, to bind, to co-immunoprecipitate, ...) was compiled and used to filter the PAS. The resulting set contained 2135 potential non-interactions. Again, a sample of the results was inspected manually. Using the confidence score described in the next section, the top, median and bottom 20 sentences were evaluated. This resulted in accurracy values of 95%, 45% and 15% respectively.

Considering that all freely-availabe articles from MEDLINE where analyzed, one has to state that this number of potential non-interactions is very low. This finding might be explained by the fact that negative results are less likely to be mentioned in biological publications but might also point towards a recall problem of Excerbt. For manual annotation of the Negatome, however, this result provided a manageable, yet sufficently large corpus.

Examining the results showed that the main remaining source of error was the ambiguity of protein symbols and compounds. Since the same names were often used for both kinds of entities, a lot of non-interactions between two compounds or a compound and a protein were detected as protein-protein non-interactions. To overcome this problem we applied different disambiguation algorithms and heuristics, including the WSD approach presented in the previous chapter. Unfortunately, none of the applied methods could significantly improve the situation. This was most likely due to the fact that all the applied methods worked on a sentence level but the ambiguity often could not possibly be resolved on such a level. Often there were the same formulations used to describe non-interacting compounds or proteins. A more successful disambiguation method would need to target the passage within the paper where the relevant entities were introduced.

Based on the unsuccessful attempt to disambiguate proteins and compounds we decided to leave the compounds in the data set and annotate them along with the protein-protein non-interactions. Furthermore, a second data set was created containing protein-protein, protein-compound and compound-compound non-interactions in case the non-interaction containing compounds would be of interest

in the future. The resulting data set consisted of 4177 potential non-interactions. Again the quality of the data was checked by looking at samples. Accurracy values of 90%, 65% and 50% were obtained for the top, median and bottom samples respectively. As can be seen, if all three kinds of non-interactions are considered correct, the results significantly improved. This indicates that text mining systems can benefit greatly from more sophisticated disambiguation algorithms within the biomedical domain.

4.4 Confidence Score

Including a confidence score into a text mining system has different advantages. It can be used to produce results with higher precision when a confidence threshold is applied. Or if the results are validated by annotators the results can be ordered by the score, so that the most likely results are presented first. This proves especially helpful when more results are returned than could be annotated manually.

For the development of Negatome 2.0, Excerbt was extended to deliver such a score. This way an annotator could start with the more likely sentences and stop once enough non-interactions were found. The confidence score was based on simple surface features of the sentence and the predicate-argument-structure. It was defined as follows:

$$c = \alpha * (1 - \frac{l_0}{100}) + \beta * (1 - \frac{l_1}{100}) + \gamma * RT + \delta * NW + \epsilon * EQ + \zeta * (1 - \frac{l_s}{1000}) \tag{4.1}$$

Here l_0 and l_1 are the length of the arguments 0 and 1 found by Senna. Longer arguments might indicate that the sentence is more complicated or that the argument contains additional relative clauses. RT referes to the relation type. If the Excerbt relation is a binding or interaction this value is one. However additional less likely relation types were allowed but discounted by assigning a lower RT value. For the types 'functions_as' and 'expression, is_a' RT is 0.5 and for 'is_a' it is 0.25. NW corresponds to the term that Senna tagged as ARGM-NEG. If this is one of the words 'not', 'never' or 'unable' NW is one, otherwise 0.5. EQ refers to whether the hits in ARG0 and ARG1 are the same. If this is the case, most likely there was an error in the relation construction. Thus, EQ is 0 if this is the case and 1 otherwise. The relations are still kept in the result set, however, since the negated predicate and the existence of a protein in an argument role stills points towards a non-interaction, just not the one predicted by Excerbt. Finally, l_s stands for the length of the sentence. Shorter sentences are easier to analyze and thus more likely to contain a true non-interaction. α, β, γ, δ, ϵ and ζ are weights of the different features. For the Negatome annotation an equal weighting of $\frac{1}{6}$ was applied. These weights migh still be optimized in the future.

As mentioned above the confidence score seems to correspond well to the classification precision of Excerbt. The highest scoring candidates in the sample sets performed 40 - 80 % better than the lowest scoring ones. The use of confidence scores is especially useful, when text mining results are used in conjunction with manual annotation and can significantly accelerate the annotation process.

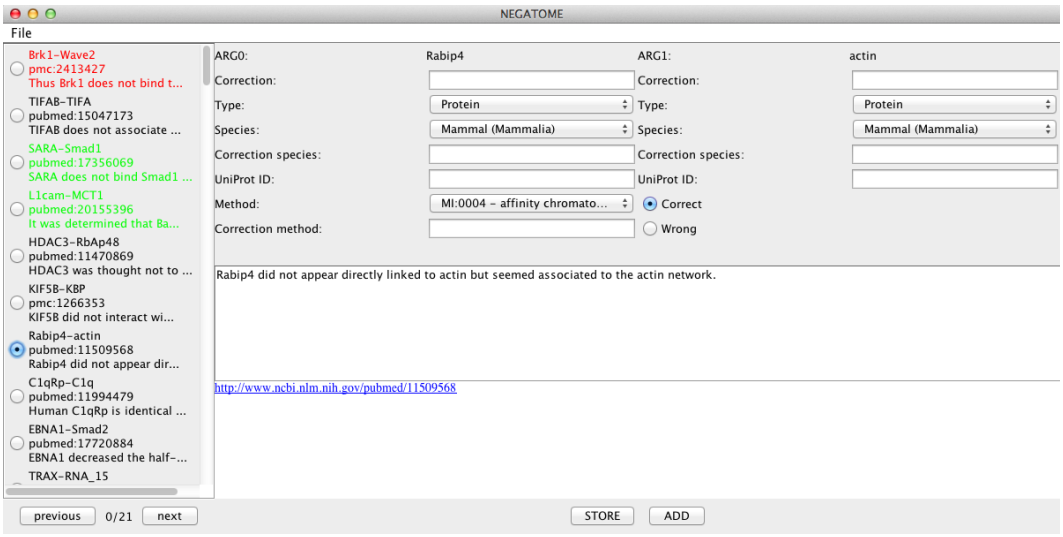


Figure 4.1: Tool for annotating the non-interactions proposed by Excerbt.

4.5 Results

As described above, the filtering approach increased the precision so that more than 50% of the non-interactions were classified correctly. The sample evaluation also pointed towards the usefulness of the confidence score. Based on these promising results all the remaining candidate non-interactions were evaluated by manual annotators to create Negatome 2.0. To facilitate this process an annotation tool was developed that was intended to enable an efficient annotation process. Figure 4.1 shows the tool.

As can be seen, ARG0 and ARG1 of the relation found by Excerbt are shown. The annotator has the possibility to enter a correction in case the arguments are incorrect or to flag the whole entry as wrong. Additionally, the annotator has the possibility to further add contextual information about the relation like the species it was observed in, the UniProt IDs of the proteins and the method that was used for the detection (using the HUPO-PSI controlled vocabulary (Kerrien et al., 2007)). To facilitate the decisions the sentence, from which the non-interaction was extracted, and a link to the corresponding publication is shown.

The 2134 potential non-interactions of Excerbt were manually evaluated using this tool. Furthermore, non-interactions from the publications proposed by Excerbt were included, if the annotators noticed them while searching for information about the propositions of Excerbt. This process resulted in 895 non-interaction between proteins and 119 between proteins and compounds. The results were largely based on mammalian proteins (86%). 64 of the non-interacting protein pairs (NIPs) include at least one splice variant. Around 90% of these were proposed by Excerbt while the rest was added from the investigated publications. The non-interactions found by text-mining-aided manual annotation were complemented with further ones derived from three-dimensional structures of PDB biological units (Kouranov et al., 2006). Combined with the results from Negatome 1.0, this resulted in a manual data set of 2171 NIPs and a structural one of 4397. Both data sets were then checked for whether a PPI

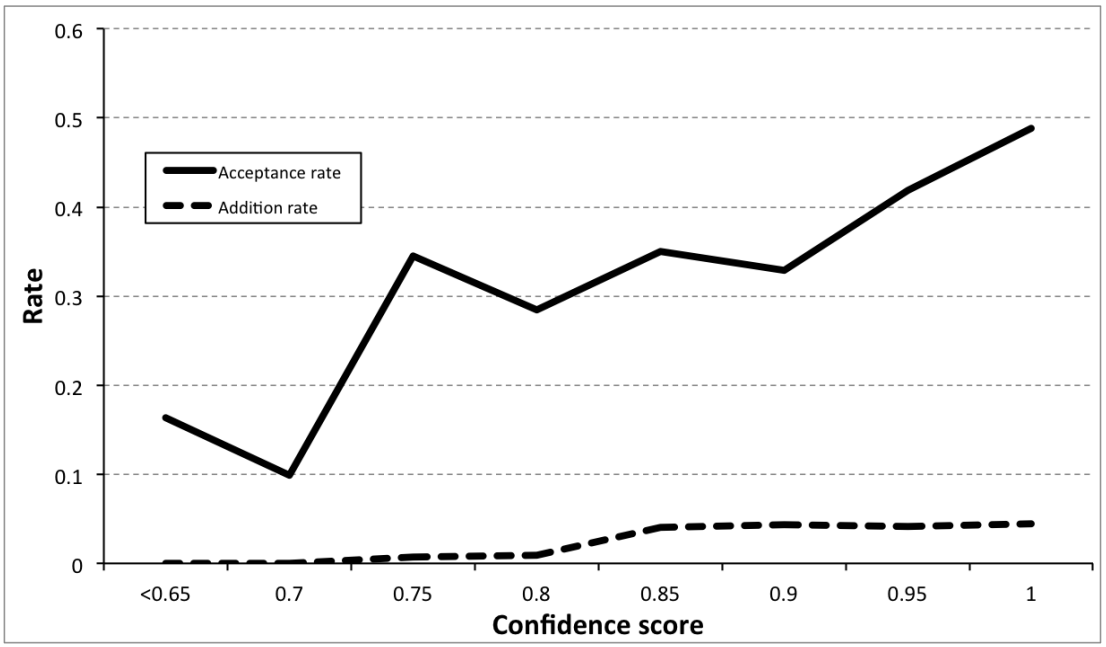


Figure 4.2: Acceptance Rate of Negatome annotation. One can see what ratio of Excerbt proposals was accepted by the manual curator and how many non-interactions the annotator added from papers proposed by Excerbt.

between them was reported in IntAct. Filtering out these, resulted in more stringent data sets of 1991 and 4161 NIPs respectively.

The ratio of correct Excerbt predictions and of additionally found non-interactions with respect to the confidence score is shown in Figure 4.2. The acceptance rate and the addition rate are plotted in relation to the confidence score. The acceptance rate corresponds to the ratio of Excerbt propositions that were included into the final version of Negatome 2.0. The values are lower than the performance measures mentioned before because of the strict acceptance criteria of the Negatome. Excerbt works on rather coarse grained meaning definitions while a more fine-grained distinction was made by the annotators of the Negatome. For example, Excerbt does not distinguish between proteins and protein complexes or between mutated and wild type proteins. In the Negatome, on the other hand, the distinctions were made and protein complexes as well as mutated proteins were excluded. Furthermore, the context of the non-interaction was considered by the annotators. For example in the sentence “Inversin does not coprecipitate with tubulin after addition of colcemid.” Excerbt correctly identifies the non-interaction between inversin and tubulin. However, Excerbt lacks a proper context resolution, which could determine that the non-interaction only exists after the addition of colcemid. For this reason, the non-interaction was not accepted by the annotators. The addition rate, on the other hand, corresponds to the amount of additions made by the annotators that were not proposed by Excerbt but that they came across while reading the papers proposed by Excerbt. It can be seen that both rates correlate with the confidence score.

In order to compare the text-mining-aided annotation results with the ones derived from an unguided annotation approach the overlap between Negatome 1.0 and Negatome 2.0 was analyzed. This analysis yielded a very small overlap of only 15 NIPs. Likewise, the overlap between structure-based and

annotation-based pairs in Negatome 2.0 was only 36 pairs. From this it can be concluded that depending on the method of information gathering/creation very different results are obtained. For example the structure-based analysis created de novo predictions while the annotation could only deliver results already described in the literature. In order to obtain an unbiased and comprehensive resource it therefore seems recommendable to combine multiple approaches.

Because of the low observed overlap between the annotation results, the text mining results were analyzed further. Here, the reasons for not finding the NIPs described in Negatome 1.0 were at the focus of attention. A manual error analysis was performed on a subset of 40 randomly chosen NIPs from Negatome 1.0. Only one of the NIPs was correctly identified by Excerbt. The availability of information turned out to be the biggest problem. 32 of the NIPs were contained either in the full text of publications where Excerbt could only access the abstract, or in figures or tables. This points towards the larger problem that many biomedical publications are not freely accessible for text mining systems. Apart from this, however, the approach of Excerbt is flawed since it is not able to detect biological events that are described by elipses, anaphora or nominalizations. These linguistic phenomena accounted for four cases. Finally, in one case there was an error in Negatome 1.0, in one case an unofficial name not covered by the Excerbt ontology was used an in another case Senna made a mistake. This points towards an immense recall problem of text mining in general and Excerbt in particular.

In order to investigate the recall of Excerbt more specifically, twelve additional examples from text reachable by Excerbt were chosen. Together with the eight examples from above more meaningful explanations for the recall problems should have been detected. From these 20 NIPs still one was an error in Negatome 1.0, three were found by Excerbt, in two cases Excerbt was wrong, and in four cases unofficial names were used. The remaining 10 examples were not findable with Excerbt's aproach. Here, in one case a 'failed to'-formulation was used which could not be resolved since Excerbt has no way of interpreting nested predicate-argument-structures. Two cases were ellipses. In an ellipsis the verb is omitted since it can be derived from the context (e.g. in formulations like '... activated X but not Y'). Since Excerbt only detects biological events that are described by a verb, there was no way of correctly understanding ellipses. For the same reasons two cases could not be detected where nominalizations were used to describe the NIP (e.g. formulations like 'the non-interaction of X and Y'). Finally, Excerbt lacks an anaphora resolution module, which accounted for four missed cases (e.g. formulations like 'The two proteins did not interact'). All of this resulted in a data set, which had an appropriate size to enable a reasonable update for the Negatome but which was most probably very low compared to the amount of non-interactions that were actually described in the biomedical literature.

4.6 Conclusion

The use of a text mining system to assist the manual annotation process proved useful. The developed confidence score was shown to correlate with the quality measures of Excerbt. The precision of the approach was sufficiently high for this application. However, different problems of data availability and the Excerbt approach restricted the number of results making it questionable whether this method is applicable to more specific problems. Further contextualizations from PAS would seem very useful to provide annotators with additional information (e.g. species, used method) in order to design more precise filters and to further facilitate the annotation process.

4.7 Related Work

An overview of related work concerning negation and other PAS contextualizations is given in Table 4.1.

Table 4.1: PAS Contextualization: Related work

Authors	Year	Contextualization	Domain
Sanchez-Graillet and Poesio (2007)	2007	Negation	Biomed.
Morante and Sporleder (2010)	2010	Negation, speculation	General
Agarwal and Yu (2010)	2010	Negation	Biomed.
Gerner et al. (2012)	2012	Negation, speculation, species, anatom. location	Biomed.
Sarafraz and Nenadic (2010)	2010	Negation	Biomed.
Vita et al. (2006)	2006	Experimental context	Biomed.
Wei and Collier (2011)	2011	Species	Biomed.
Rinaldi et al. (2008)	2008	Experimental method	Biomed.

5

Section Contextualization

Many documents are structured in sections or chapters. This helps to order the train of thought of the author as well as the reader. Moreover, such an organization provides the text with a certain kind of structure. Since the text within each section, however, is freely written again, such documents are called semi-structured. When analyzing the unstructured text within such sections, it makes sense to use the structure of the document as contextual information. This chapter gives an overview of typical information that can be used in semi-structured texts and presents an application where such a section contextualization was used within the course of this thesis. Since this application focussed on text mining of rare disease profiles, prior to the description of the approach an introduction to rare diseases is given.

5.1 Section Information

Every document that follows a standardized scheme at some point might be referred to as a semi-structured document. Examples of such documents are websites, medical records and scientific publications. In a HTML-website at least the headings are standardized. Furthermore, there exist standardized ways to emphasize text, link to other pages and create tables. In medical records, usually there are standardized formulations to describe the state of the patient. This can complicate the analysis because the formulations might not be grammatical anymore turning typical sentence analysis tools useless. On the other hand, the consistent formulations might make it easier to formulate a set rules for information extraction. In scientific publications, the respective journals or conferences often require the authors to follow a rigid organization of the paper. Typical sections, like 'Introduction', 'Materials & Methods', 'Related Work', 'Results', 'Discussion' and 'Conclusion', are frequently present and hint towards the contents of the respective text.

Standardized formulations, a standardized organization in sections or additional annotations in the text (e.g. links in a website) can be used in various ways. The practical realizations, however, are manifold and diverse. There does not exist a standardized way of dealing with semi-structured documents. Instead, the many different ways to structure documents give rise to a variety of approaches to make use of them. Many of these are closely linked to the respective domain and problem at hand. Yet, at least a generalized categorization of these approaches and some examples of these should be given in this section. The structured information can either be extracted directly (e.g. to detect cross-references between scientific publications), or used in order to benefit other text mining approaches. In the latter case, the use of structured information can be subdivided in approaches that use it to produce additional features for the respective text mining algorithms and those that use it for filtering. Thus, the structure of the document can be made use of by creating specialized features for classification algorithms or by restricting the search space for certain information.

If the text consists of common sections or chapters, these divide the whole text in different zones that belong to different categories. This additional information can be fed directly into information extraction machine learning algorithms to improve their performance (Chieu and Ng, 2002). So, if e.g. a hedge detection algorithm (an algorithm that tries to detect negations and speculations) is given an additional feature of whether the extracted event is from the discussion section or not, it might be easier for it to determine whether the event is a fact or speculation, since specualations should be more common in the discussion section. Likewise, if e.g. a text categorization algorithm is trained to distinguish between different kinds of scientific publications, it might be helpful to know whether the text contains a 'Materials & Methods' or a 'Results' section which typically exist in research but not in review articles.

A related strategy of how to use structured elements can be found when instead of machine learning algorithms context free grammars are used for information extraction. This might be useful, when the structured elements are not section headings but are part of the text, like e.g. HTML-tags in websites. Here, the structured elements can easily be defined as elements of the grammar. This is a very natural way of integrating structured and unstructured data. Furthermore, context free grammars have the advantage of being able to model dependencies reaching over several tokens which certain machine learning algorithms like conditional Markov chain models can not (Viola and Narasimhan, 2005). A similar strategy can also be implemented in other rule based systems besides context free grammars (see e.g. (Muslea et al., 1998)).

Furthermore, there might be keywords within the text that bear special meaning and could thus be treated like structured elements. For instance, in a document announcing a seminar, the lecturer and the place of the seminar might be mentioned after certain standardized formulations like 'Who: ...' or 'Where: ...'. Using a priori knowledge about such a structure in the document can significantly improve the information extraction results (Chieu and Ng, 2002).

Apart from feature and rule generation, the zones of the document can be used for filtering. Here, certain information extraction searches are restricted to certain sections (Smith et al., 1997). For example, if one is interested in extracting the analysis method of an experimental paper, one could restrict the search to the Materials & Methods section. This way the false positive rate that might occur by extracting methods mentioned when describing related works of other authors might be reduced. Such techniques make use of Grice's maxim of relevance mentioned in section 2.2. Since the topic of a

certain section is predetermined by the structure of the document, the author of the text is inclined to restrict his remarks to this topic and not to focus on it in other parts of the text.

In addition to that, strict structural rules can simplify event extraction. Take, for instance, the seminar announcement example from above. If the location $L1$ and the lecturer $L2$ are extracted from the text, one can immediately deduce that these information bits belong to the seminar S described in the heading or elsewhere in the document. Thus, the relations *located_at*$(S, L1)$ and *taught_by*$(S, L2)$ can be formed, even though the relations are not described in any sentence and thus the entities are not connected by any syntactic structure. Instead the structure of the document provides the required additional information to connect the pieces. One refers to the task of extracting information from documents like this as single-slot information extraction. In such documents, event extraction is simplified by the fact that one argument is always given by the topic of the text. The opposite, where multiple topics might be discussed in a document, is referred to by the term multi-slot information extraction (Chieu and Ng, 2002). The same distinction can be made for single sections. The author field in a scientific paper is for example such a single-slot section. Each term given there must be a researcher and stands in a *published* relation to the publication itself.

An additional method that can utilize certain types of structured or semi-structured information is bootstrapping. Bootstrapping tackles the problem that information extraction typically requires large amounts of manually annotated data. This data can either be in the form of annotated training examples or lexical/ conceptual resources. Either way, huge manual effort is necessary. Bootstrapping tries to reduce this by first learning a simple information extraction model and iteratively extending it by learning from the results obtained in the previous iteration (Maedche et al., 2003).

For example, large ontologies can be learnt from small seed ontologies by bootstrapping. In such a case, patterns are learnt that typically describe the elements in the different classes of the ontology and these patterns in turn are used to further fill the ontology, which in turn leads to the extraction of more patterns in the next iteration. Borrowing an example from Califf and Mooney (Califf and Mooney, 1999), the utterances:

'located in Atlanta, Georgia'

and

'offices in Kansas City, Missouri'

could be generalized to a pattern of the form

'in <POS tag = NNP, max. length = 2, ontology category = city> , <POS tag = NNP, ontology category = state>'.

Here, the expressions in the angle brackets stand for one term that fulfills the given constraints. Such a pattern can then be used to learn new cities or states for the ontology and a larger ontology, in turn, can be useful in finding additional patterns.

Bootstrapping is inspired by the language acquisition process of children. Researchers like Lila Gleitman (Gleitman, 1990) and Steven Pinker (Pinker, 1994) analyzed the way children learnt new word meanings and proposed a bootstrapping strategy where new words are learnt from the syntactical and semantical constraints imposed on them by the known words in the context in which they are used. Exemplarily,

if a child tries to learn the meaning of the word 'glip' and hears it used in sentences like 'I glipped the book' and 'I glipped the book from across the room' but never in sentences like 'Glip that the book is on the table!' and 'What John did was glip the book', it can draw several conclusion. The first two sentences imply that 'glip' is something that can have a direction and something that can be done to an object. The other two would suggest that glip is something voluntary like an action. But since the word is never used in such sentences it can be deducted that 'glip' is involuntarily and not an action. Combining these bits of information the child can then conclude that 'glip', as an nonvoluntary non-action that can have a direction and be applied to an object, might mean something like 'see', which possess the same features (Pinker, 1994).

Such methods build on patterns that need to be as rigid as possible. Therefore, semi-structured or even structured documents are ideal for applying bootstrapping techniques. The more structured the utterances are, the smaller is the danger of a so-called semantic drift. This drift describes the viscious circle where wrong terms or patterns are learnt which in turn lead to more mistakes, since the learnt knowledge is used as basis in further iterations. Consequently, bootstrapping has been frequently used on semi-structured documents (see e.g. (Carlson et al., 2010a; Carlson and Schafer, 2008)).

5.2 Rare Diseases

Traditionally, the main focus of research has always been on diseases that affect many people. Here, the benefit from providing cure as well as the financial incentives were the highest. Besides, the commonly researched diseases, however, there exists a vast amount of rare diseases. Each of these diseases affects only comparatively few individuals. Yet, because of the large amount of rare diseases, there are many millions of people suffering from one of these. For this reason, since the 1970s rare diseases slowly moved in the focus of research, resulting in different legislation for the development of so-called 'orphan drugs' for rare diseases since the 1980s (Bavisetty S, 2013).

A rare disease is defined by its prevalence. The threshold, however, differs regionally. In the United States, a disease is considered rare if less than 200,000 people in the USA are affected by it. In Europe, the threshold is at less than one in 2,000 affected Europeans (HHS, 1989). These definitions result in over 6800 rare diseases which affect ca. 25 million people in the US and ca. 30 million in Europe. This constitutes around 8% of the population. It is estimated that 80% of rare diseases are genetically determined (Bavisetty S, 2013; Eurodis, 2005; HHS, 1989).

The development of treatments for rare diseases is difficult. It is complicated by a lack of knowledge about the disease, a lack of patients to investigate and to use for clinical trials of potential drugs and the decreased market potential of the drugs. While the amount of patients suffering from a rare disease is not amenable to influence, there has been some effort to overcome the lack of knowledge and financial appeal. Legislation both in the US and Europe created incentives for research of rare diseases. As a consequence of this, research on different rare diseases increased (Wästfelt et al., 2006). Furthermore, knowledge bases for rare diseases were created that try to collect the comparatively sparse information about them.

Most notably Orphanet (Orphanet, 2014) is a portal for rare diseases and orphan drugs. It provides basic information about currently 6,760 rare diseases. Furthermore, Orphanet offers information on

the prevalence of diseases, orphan drugs and research infrastructures in Europe. The Office of Rare Diseases Research (ORDR) (ORDR, 2014) at the National Center for Advancing Translational Services might be considered the American counterpart to Orphanet. It also provides a list of rare diseases and provides certain basic information for them. Additionally, it links to Orphanet and other relevant resources. Apart from that, the rare diseases database of the Swedish National Board of Health and Welfare (Greek-Winald et al., 2010) should be mentioned. It contains more detailed descriptions of over 300 rare diseases. Besides these specialized resources, general disease resources play an important role in collecting information about rare diseases. Since around 80% of rare diseases are genetic diseases, especially resources that focus on genetic aspects, like GeneReviews (Pagon RA, Adam MP, Bird TD, et al., 2014) or OMIM (OMIM, 2014), are of interest.

The low prevalence of rare diseases poses a major problem for physicians when trying to diagnose patients. Since the physician hardly ever gets to see a patient with a rare disease, his knowledge and experience of the respective disease is likely to be very restricted and the danger of a misdiagnosis increases. Knowledge resources, like the ones described above, could build the foundation for tools that support physicians in their decision making. Here, decision support tools could assist the physician by proposing common and rare diseases that match the symptoms of his patient.

5.3 Single-slot symptom extraction for a decision support tool

In order to use existing knowledge bases for decision support tools, the knowledge in them needs to be in structured form. Unfortunately, Orphanet and the ORDR only provide little information and the information in GeneReviews and the rare diseases database of the Swedish National Board of Health and Welfare is only available in semi-structured articles. In order to overcome this, in the course of this work text mining was used to extract symptom-disease relations from GeneReviews. The extracted knowledge was stored in a new rare diseases database called PhenoDis. Furthermore, a website to manually check and extend the text mining results and to present the results to rare disease researchers was created. The data in the database was used to create a decision support tool. This work was done in collaboration with Andreas Ruepp and the annotation group at the Institute of Bioinformatics and Systems Biology (IBIS) at the Helmholtz Center. Additionally, Jon-Magnus Meier (created the database and website) (Meier, 2014) and Maximilian Herzog (parsed OMIM and OrphaNet and mapped the entities) (Herzog, 2014) worked on this project under my supervision in the course of their Bachelor Theses. I designed and implemented the text mining approach, chose the significance weighting scheme and the features for the symptom mappings, designed and implemented the N-gram analysis and the decision support tool, and evaluated the system.

The biological experts I collaborated with chose GeneReviews as the most qualified resource for the description of rare diseases. GeneReviews consists of 599 chapters that describe diseases or give overviews over a collection of related diseases. The different chapters follow a somewhat rigid scheme depending on their type. Overview chapters typically consist of the sections 'Summary', 'Definition', 'Causes', 'Evaluation Strategy', 'Genetic Counseling', 'Resources', 'References' and 'Chapter Notes'. Furthermore, some chapters like 'Management' and 'Molecular Genetics' are optional. Disease chapters consist of the sections 'Summary', 'Diagnosis', 'Clinical Description', 'Differential Diagnosis', 'Management', 'Genetic Counseling', 'Resources', 'Molecular Genetics', 'References' and 'Chapter Notes',

Figure 5.1: Simplified version of an information extraction pipeline that can be used for relation extraction in single-slot tasks.

with no optional sections. We decided to focus on the disease chapters and examined the sections for where the disease-symptom relations were described.

This manual examination showed that the symptoms were largely described in the 'Clinical Description' section and furthermore that this section was nearly exclusively dedicated to symptoms of the respective disease. Based on these observations the information extraction task at hand could be classified as a single-slot IE task. As described above, this reduced the complexity of the text mining problem. Since all symptoms described in this section belong to the disease of the respective chapter, the relation extraction task could be reduced to a named entity recognition task. Thus, all symptoms found in this section form a symptom-disease relation with the disease that is indicated in the heading. This simplified the typical text mining pipeline as it was shown in Figure 1.7 to the shorter pipeline shown in Figure 5.1.

Like in Excerbt, the named entity recognition was performed with a dictionary-based approach. However, in order to have a proper mapping to the resources instead of the Excerbt vocabulary other established lexical resources were used. For this MedDRA, the Human Phenotype Ontology, the Mammalian Phenotype Ontology, ICD10 and a self-constructed N-gram-based dictionary (see section 7.2) were tested for their applicability. The results of this evaluation are described in the following section.

Based on the results of the single-slot relation extraction, a decision support tool was developed. In addtition to the text mining results, structured information from Orphanet and OMIM was integrated into the database. The information about which symptoms are present or not for a specific disease were collected in so-called disease profiles. These profiles consisted of a binary symptom vector that contained a 1 at every position belonging to a certain symptom if this was present in the disease and a 0 otherwise. Furthermore, for each symptom its significance was calculated using the following formula:

$$w_i = \frac{o_i}{\max_j o_j} \tag{5.1}$$

The significance was used as a weight in the decision support tool. By defining each weight w_i in this way the significance are normalized by the ratio between the occurrences o_i of symptom i and the number of occurrences of the most frequent symptom. Besides this, a normalization by the total amount of symptom occurrences was tested but delivered worse results due to the fact that the total number of occurrences was way larger than the maximum number of a single symptom and thus the weighting effects were diminished.

The decision support tool was based on calculating similarities between disease profiles. In this work, the cosine similarity was used in this step. This similarity is defined as follows:

$$s = \frac{\sum (X_i * Y_i)}{\sqrt{\sum X_i^2 * \sum Y_i^2}} \tag{5.2}$$

Here, the similarity s between two vectors X and Y is calculated by calculating the cosine of the angle between the two vectors. The similarity measure was used both with the binary and the weighted symptom vectors. It was evaluated on a dataset based on the rare disease database of the Swedish National Board of Health and Welfare (Greek-Winald et al., 2010). The results are shown in the following section.

In addition to the NER results, disease-symptom relations taken directly from OMIM and Orphanet were included in the database. Consequently, the mentioned symptoms were also included in the vocabulary. As far as there existed mappings the diseases and symptoms mentioned in the different resources were consolidated. In addition to that a String similarity algorithm was used to map additional entities. This algorithm changed the form of a String in the following ways:

- the Strings were transformed to lower case
- the whitespace surrounding the String was removed
- a range of special characters was removed (commas, brackets, consecutive white space characters, quotation marks, ...)
- expressions between brackets were removed
- the words were replaced by their lemmas
- for terms with up to five words all permutations of the words were used for comparison

The so mapped diseases and symptoms were combined with the text mining results and stored in the PhenoDis database.

5.4 Results

A dictionary-based named entity recognition approach is highly dependent on the quality of its dictionary. For this reason, the following four different dictionaries were tested for their applicability: the Human Phenotype Ontology (HPO) (Robinson et al., 2008), the Mammalian Phenotype Ontology (MPO) (Smith et al., 2004), ICD10 (World Health Organization, 2014), and MedDRA (MedDRA, 2014). Beside these established vocabularies a N-gram analysis was performed in order to create an additional alternative. Since such an analysis is a form of using corpus information to derive a dictionary, it is explained in the corpus contextualization chapter (see section 7.2).

The N-gram analysis was performed on the Clinical Description section of all disease chapters of GeneReviews. In total 14.407 N-grams were extracted. All N-grams that occured at least seven times (693 N-grams) were manually checked. The process confirmed the observation that symptom terms were the dominant category in the Clinical Description sections. Apart from that, also terms describing age, anatomical regions and methods could be determined as recurring and were thus also extracted.

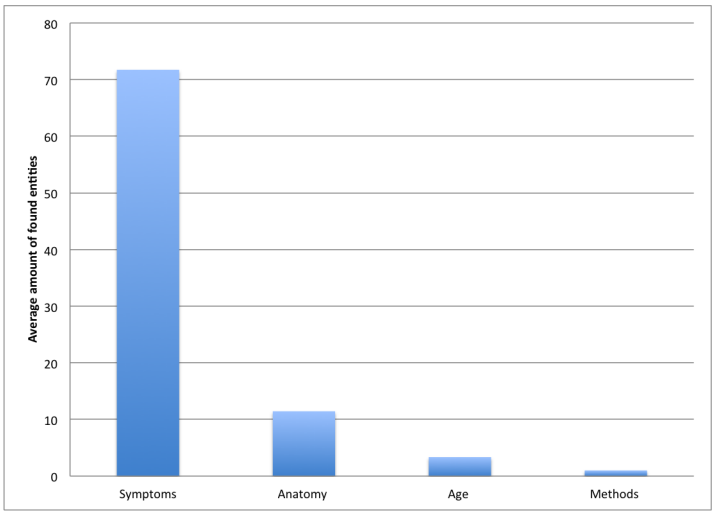


Figure 5.2: Average amount of entities found with NER using all vocabularies.

The dictionary dervied from this n-gram analysis consisted of 162 symptoms, 105 anatomical terms, 15 method terms and 8 age terms. Figure 5.2 shows how many terms from all vocabularies were found on average in one Clinical Description section. This figure does not contain any consolidation procedures. Thus, duplicates from multiple vocabularies could be present. Figure 5.3 shows how many symptoms were found on average from each of the single vocabularies.

As can be seen, MedDRA turned out to be the most suitable vocabulary with 37.8 symptoms on average. ICD10 on the other with only 0.32 symptoms hardly found anything at all. The N-gram analysis found more than 6 symptoms on average which was a quite high turn-out considering that the vocabulary consisted of only 162 terms in comparison to the thousands of terms from the other vocabularies. The low turn-out of ICD10 can be explained by the fact that this vocabulary was written as a reference for humans and contains formulations like the following that are unsuitable for text mining without further processing because they would not occur in texts like that:

> Epidemic louse-borne typhus fever due to Rickettsia prowazekii
> Spotted fever, unspecified
> Acute paralytic poliomyelitis, other and unspecified

Based on these evaluation results, my collaborating biologists chose MedDRA and HPO as vocabulary for the decision support tool. Even though MPO showed slightly higher NER values, HPO was chosen because of its specific focus on humans. In addition to this evaluation of the amount of found entities, the different NERs were also manually evaluated on a sample of five GeneReview chapters chosen by the biological experts. These chapters contained 417 symptom mentions. The results of this evaluation can be seen in Table 5.1. The ICD-10 vocabulary was left out of this evaluation, since the ICD-10-based NER did not find any term within the sample texts.

As expected in a dictionary-based NER approach, the precision values are very high. Among the remaining errors were enumerations that broke the split off part of the term or symptoms described in whole sentences. In addition to that, some broad superior terms (e.g. all from the HPO) created the

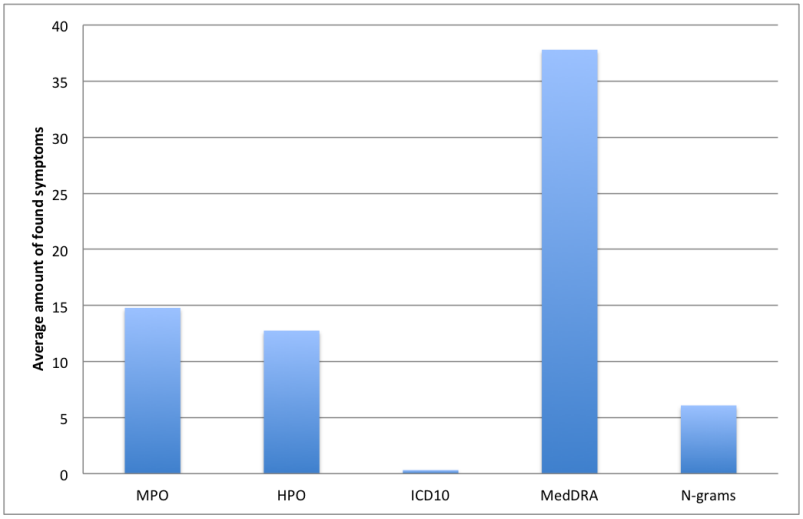


Figure 5.3: Average amount of symptoms found with NER using different vocabularies.

Table 5.1: NER Performance on sample of GeneReviews with different vocabularies.

Dictionary	Precision	Recall	F-Measure
HPO	0.80	0.17	0.28
MedDRA	0.85	0.58	0.69
MPO	0.96	0.38	0.54
Ngram	0.88	0.19	0.31

largest amount of false positives. To avoid this, in future applications a blacklist ignoring these terms created by biological experts would prove valuable. The main difference of the dictionaries lies in their recall. Here, MedDRA showed by far the best performance. Consequently, MedDRA was chosen as the main reference for symptoms for the whole project. The symptoms from HPO and the structured disease-symptom resources OMIM and Orphanet were mapped on MedDRA terms where possible.

The NER based on all these vocabularies resulted in disease profiles containing 44.1 symptoms on average. These profiles were used in the decision support tool. The quality of the decision support tool was evaluated on the Swedish Rare Diseases Database (SRDD) . From the 175 disease descriptions in SRDD 60 could be matched to diseases described in GeneReviews using an exact string match (here the String similarity matching was not used in order to guarantee the identity of the diseases). From these 60 disease descriptions the 'Symptoms'-sections were extracted. From each of these sections disease profiles were derived using the NER and for each of these 60 profiles the similarity to each of the 535 GeneReviews disease profiles was calculated. In each case the GeneReviews were ordered according to their similarity and the position for the matched disease profile was determined. The results of this evaluation can be seen in Figure 5.4 and Table 5.2.

As can be seen, the decision support system showed very promising results. In over half the cases the correct disease profile had the highest similarity score. In 75% of the cases it is in the top 5 results (top 1%) and in 95% in the top 26 (top 5%). The significance weighting helped to improve the results. The

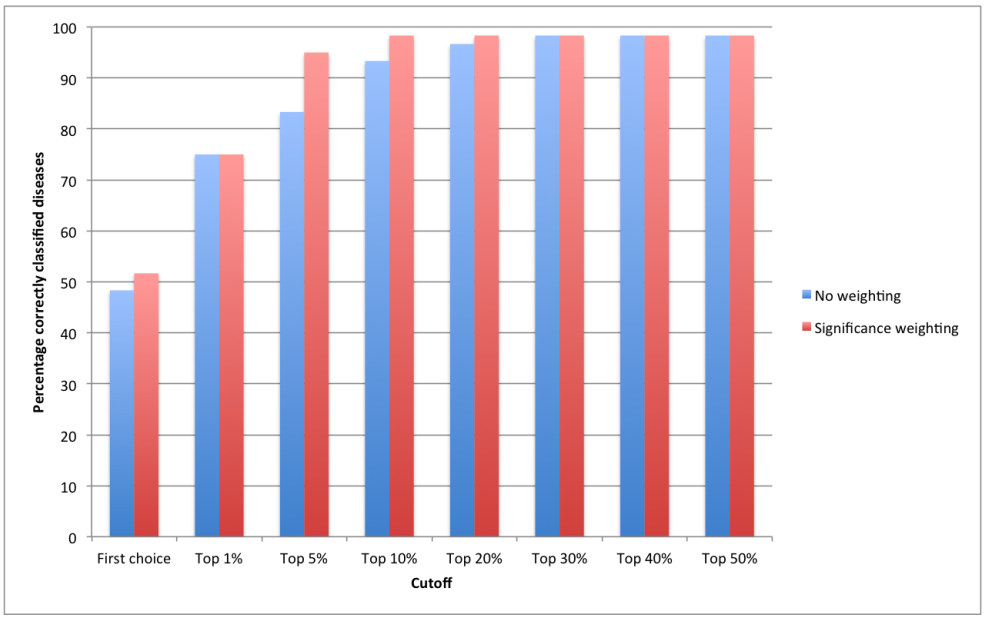


Figure 5.4: Performance of the decision support tool with and without weighting for different cutoffs relative to the total amount of diseases.

Table 5.2: Evaluation of decision support tool.

Approach	Completely correct	Average rank	Median rank
No weighting	48.3%	17.48	2
Significance weighting	51.7%	12.15	1

average rank of the correct disease profile was 12.15 compared to 17.48 without the weighting and the median was 1 compared to 2. Using the weighting scheme all but one of the test cases where within the the top 10% of the similarity ranked diseases.

5.5 Conclusion

The decision support tool performed very well even though the simplified NLP pipeline was used and the vocabulary was still completely automatically merged without any blacklist or manual consolidation. Likewise, the similarity ranking algorithm might be replaced by more sophisticated reasoning or machine learning algorithms. Furhter work on these issues may create an extremely valuable tool to physicians to support them in their daily work and to prevent misclassifications of rare diseases. And even the current state of the decision support tool has a remarkable precision.

More generally, this application also showed the potential of using section contextualization for text mining. Based on these very good results a contextualization similar to the one described here might prove to provide valuable context information also for general event extraction systems.

5.6 Related Work

Table 5.3 gives a brief overview of some of the most important approaches related to the decision support tool.

Table 5.3: Decision Support Systems: Related work

Authors	Year	Approach	Domains
Graber and Mathew (2008)	2008	NLP, rules	Pediatrics
Segal (2004)	2004	Pattern matching	Medical
Köhler et al. (2009)	2009	Semantic similarity metrics	Medical
Dragusin et al. (2013)	2013	Information retrieval	Rare diseases
Yan et al. (2006)	2006	Artificial neural net	Heart diseases
Stylios et al. (2008)	2008	Fuzzy cognitive maps	Language pathology, speech pathology, obstetrics
Kuperman et al. (1991)	1991	Rule-based	Medical
Myers (1987)	1987	Scoring scheme	Internal medicine
Middleton et al. (1990)	1990	Belief networks	Internal medicine
London (1998)	1998	Bayesian reasoning	Medical
Coulson et al. (2001)	2001	Rule-based	Genetic risk assessment
Wells et al. (2007)	2007	Rule-based	Cardiovascular diseases

6

Text Contextualization

Text contextualization can work in the same way as the section contextualization described in the previous chapter. The type of text might influence the information extraction methods performed on it. Apart from that, the text seems to be a quite natural unit for contextualization. Because of this, several ways of text contextualization exist. This chapter is intended to give an overview of these. Here, especially anaphora resolution, for which an algorithm was implemented in the course of this thesis, will be presented in detail.

6.1 Text Information

The main focus of text mining is the extraction of relations and susequently events. Besides the relations it contains, a text can, however, have certain features on its own. It can have a topic, a type (e.g. a research paper or a review) or can be written in a certain style. Different techniques exist to capture some of these aspects.

Topic detection identifies the topic of the text. It is often accompanied by approaches to track the identified topics over a stream of documents (Allan et al., 2005). This research is mainly focussed on classifying news articles or more recently social networking sites like Twitter (Cataldi et al., 2010) but can potentially also have value in other domains and as part of other problems. E.g, topic detection can be used as part of a summarization system or to distinguish text types.

Text categorization techniques can be used to distinguish text types or more generally to order a collection of texts in different categories. Depending on whether these categories are predefined or not, the approaches to text categorization differ. After early rule-based expert systems, statistical and machine learning approaches got established. If there exist predefined categories, supervised methods can be trained to assign the text in question to these categories. Furthermore, unsupervised approaches

can cluster documents without the need to define categories a priori or arrange documents in a high dimensional vector space that represents their semantic relatedness (Sebastiani, 2005).

Besides these classification tasks, a variety of text measures have been developed. Sentiment analysis captures the mood or opinion transported in a text. Here, efforts exist to predict psychological diseases based on texts written by potentially ill people (Wald et al., 2012) or to associate sentiment values with stock market prices (Bollen et al., 2011). Additionally, there exist formulas that try to determine the readibility of a text. An example of a text measure comparable to the readibility formulas was developed in the course of this work. A description of it can be found in Appendix E.

All of these methods that take the complete textual information into consideration can complement and enhance existing text mining systems or be used to solve problems on their own. In information retrieval, they can be used as filters to restrict the collection of considered documents by topic, document type or readibility. In this connection, especially the text measures can also be used to influence the ordering of returned results in information retrieval systems. For event extraction, these text analysis approaches can provide valuable context information. It might e.g. make a difference, if a fact was extracted from an article that mainly talks about a completely different topic or one that focusses on the issue. Likewise, the type of article and the language quality of the text could influence the reliability of an extracted event. Different text types often come with different structures. Here, approaches similar to the ones described in the previous chapter could be applied, if these text types come along with a certain type of structure.

Furthermore, text-oriented linguistic disciplines have to be considered when describing approaches to extract relevant information from texts. Pragmatics, discourse analysis, and text linguistics have already been described in chapter 2. Pragmatic approaches that try to interpret an utterance in the context of the text it occurred in can be used to improve information extraction. Discourse analysis and text linguistics, in turn, can be used to extract additional events. For example, by analyzing semantic connections that are not explicitly instantiated, one can extract additional information that would be lost if one sticks to the sentence level in its analysis. Picking up the example from Bliss from Katherine Mansfield from section 2.3: “He wants to write a play for me. One act. One man. Decides to commit suicide.” Using the pragmatic analysis of the coherence of the text, one could deduce that the man from the third sentence is meant to be the subject of the fourth sentence. This information is unattainable when one only looks at sentences.

In addition to that, pragmatic approaches for resolution of abbreviations and anaphoras can support a text mining system to find more events. The problem of the frequent use of abbreviations was already pointed to in chapters 1 and 2. There the focus was mostly on distinguishing them from other abbreviations and normal short words. This could be supplemented by the problem of abbreviations that are newly introduced in texts. An analysis on the basis of sentences would only be able to detect the proper entity in the sentence where the abbreviation is introduced but not in the remaining of the document. Instead a text-based analysis, that keeps the association between entity and abbreviation in its working memory while processing the sentences of the text, would be required. Resolving abbreviations consists of two tasks: the identification of an abbreviation definition and the redirection of the mentions of the abbreviation to the referenced entity. Since the latter is rather trivial, researchers commonly focus on approaches for the first task. Rule-based approaches for abbreviation resolution can achieve precision and recall values of as high as 98% and 97% respectively on certain data sets (Gooch,

2012). The values can, however, vary strongly depending on the used evaluation data. Additionally, machine learning approaches to abbreviation resolution exist (Zweigenbaum et al., 2013).

Finally, anaphora resolution is an important component of every comprehensive text mining system. An anaphor is a referential utterance like a pronoun that can only be understood by considering the context it occurred in. The problem of such expressions was already mentioned in section 2.4 under the name deixis. While deixis and anaphora refer to the process of referring to something in the context, an anaphor is the referrential utterance itself. Thus, an anphor is a deitic utterance. Each anphor goes along with a so-called antecedent, which is the utterance it is referring to. Multiple anphoras can refer to the same antecedent, e.g. in an expression like 'Cake is great. It tastes good and it gives you diabetes.' the two 'it' both refer to the 'cake'. Such an anaphora is called a chain. If there is only one anaphor it is called a pair (Zheng et al., 2011).

Approaches to anaphora resolution can be categorized in three classes: rule-based, supervised and unsupervised approaches. In case of the rule-based approaches, often a variety of constraints is defined to restrict the choices of a possible antecedent. Unsupervised approaches are more rare and often far less powerful than the other approaches. However, attempts to combine them with rule-based ones exist (Zheng et al., 2011). In this chapter, an anaphora resolution system will be presented that implements a constraint-based approach. It was partly implemented by Tobias Lutzenberger under my supervision during his work on his Bachelor Thesis (Lutzenberger, 2014). I designed the anaphora resolution approach, implemented parts of the system, evaluated it and performed the error analysis. The choice for a rule-based approach was based on the lack of comprehensive corpora in the biomedical domain a supervised approach could be trained on.

6.2 Constraint-based anaphora resolution

A comprehensive anaphora resolution system cannot exist on its own. Instead it needs to be integrated with a sentence analysis framework, from which it can receive the syntactic and semantic information needed for either feature or constraint generation. The constraint-based approach presented in this section is integrated with IntegreSSA, the prototype of a supersemantic event detection framework presented in chapter 10. IntegreSSA, in turn, is based on the natural language processing framework of Clueda[6].

The resolution of anaphoras is a two-step-process. First the anaphors that trigger the resolution process have to be detected and secondly the correct antecedent needs to be found. In this work, the detection of anaphors was realized by a dictionary approach. The lexical resource used for this was the newly created biological ontology described in more detail in chapter 10. Ontologies are typically hierarchical and the individuals they contain are distinguished between classes and individuals. Individuals are the leafs in the hierarchy that do not contain any children. These are concrete entities that exist in reality, like p53, Alzheimer's disease or concrete verbs that describe events like 'to activate'. Classes, on the other hand, are all the superordinate elements that have children. These are umbrella terms, like protein, disease or positive regulation, that have many different instantiations.

[6]http://www.clueda.com

Besides pronouns, these class terms are the utterances that can act as anaphors. Expressions like ’these proteins’, ’the disease’ or ’such elements’ need to be resolved to their corresponding antecedents in order to extract meaningful events. For this reason the used ontology was extended to include all relevant anaphor terms from the BioNLP 2013 anaphora resolution task. However, not every use of such class terms should be resolved. Take for example the following sentence:

> “Bacterias can be beneficial but in some cases they can cause a disease.”

Here, both ’bacterias’ and ’disease’ are class terms that can refer to concrete individuals. In this case, however, they should not be resolved since the statement is of general nature. In the contraint-based anaphora resolution approach presented here, the distinction of whether a class term is considered an anaphor is made based on the determiner they occur with. Only class terms occurring with definite articles will be resolved, while ones with undefinite articles are not. This way utterances like ’these proteins’ or ’the disease’ are considered while the ones in the example sentence given above are omitted.

Once the anaphors are detected, the resolution procedure is triggered. This procedure looks back several sentences of the anaphor for a potential antecedent. Here, each noun phrase is a candidate. However, only the ones fullfilling a certain list of constraints are regarded eligible. Among these eligible candidates the one closest to the anaphor is chosen. The constraints implemented in the approach are the following:

- Grammatical person - The gender and the number of the anaphor and the antecedent have to match. ’They’ should only be resolved to antecedents in plural and ’she’ only to female antecedents.
- Ontological category - Antecedents need to be children of their anaphors. For example, only proteins like p53 or Foxp3 should be considered eligible antecedents of ’this protein’.
- Event information - If an anaphor is part of an event, the corresponding antecedents need to match the category requirements of this event. For example, in a protein binding event derived from ’it bound to Foxp3’ ’it’ should only be resolved to proteins, since a protein binding event demands proteins as its arguments.

Additional to this, in future development a syntactic role constraint is planned. This is useful for examples like the following:

> “P53 activates Foxp3. However, it inhibits lrrk2.”

Here, the distance ordering would suggest Foxp3 as most likely antecedent of ’it’. However, because of the agreement of the syntactic role (here the subject), the correct antecedent would be p53. The feature was not yet included in the implementation of the approach due to a lack of these syntactic information. At the point of the development, the output format of the Clueda sentence analysis framework did not include information about which entity is the subject or object of a sentence, only the semantic roles were given.

The output of the sentence analysis framework is a linguistically enriched topic map. Topic maps are a way to describe information on a semantic level. The main components of a topic map are topics, assertions and occurrences (the TAO of topic maps). Topics represent the entities which act or is acted

upon. Assertions describe the relations between topics. For a biomedical protein-protein interaction event, e.g., the proteins that interact would be modelled as a topic while the type of interaction would be modelled as an assertion. Furthermore, occurrences exist that link to additional information about topics. In the previous example, this might be a weblink to the Entrez Gene webpage giving additional information about the corresponding proteins. Within an assertion, each topic is assigned a specific role, the so-called assertion role (Hatzigaidas et al., 2004). In the context of biomedical event extraction, this role refers to the semantic role the entity has with respect to the verb describing the event. These semantic roles are in accord with the roles of semantic role labeling described in section 1.3.

One important occurrence provided by the linguistically enriched topic map is an indication of the head word of each topic. Topics consist of chunks of words. For example, in the following sentence 'Awareness in dementia' would be the topic with the Arg0-role in the 'to increase'-assertion:

> "Awareness in dementia is increasing."

An intuitive way of using such arguments to extract events would be to simply look in each of them for known entities. In this given example, however, this approach fails if dementia is known and awareness is not (how it would be expected if a biomedical dictionary is used for named entity recognition). In such a case, an event describing that dementia is increasing would be extracted. Exemplarily, this is one major error source for Excerbt that makes use of this straight-forward approach. Instead, terms within a chunk can have different roles that should be considered. The head word of a chunk is the most relevant role in this connection. It determines the syntactic type of the chunk and is the part of the topic that is relevant in event extraction. For 'Awareness in dementia' the head word would be 'Awareness'. In many other cases, the head words stands at the end of the chunk. Examples of this are chunks like 'might have increased', 'this strong Foxp3 activation' or 'mostly yellow'. In some cases (e.g. genitive constructions like 'the president of the United States'), however, the head word can be in the front. In the context of anaphora resolution, the head words of topics need to be considered when resolving anaphoras.

In order to resolve anaphors properly the previously described constraints are tested on the head words of potential antecedent chunks. The list of candidate antecedent chunks consists of all noun chunks that occurred previous to the trigger word with a horizon of currently five prior topic maps. However, one exception to this typical strategy was implemented. When formulations like 'such as' or 'including' are used, the term referred to by the anaphor will most likely occur after it. Thus, in such cases the list of candidate "antecedents" consists of the noun chunks from the same sentence that occur after the anaphor. Furthermore, in such cases the number feature has to be disabled since possibly a single example from a larger group of entities can be given.

6.3 Results

The anaphora resolution algorithm was evaluated on an anaphora resolution corpus derived from the BioNLP 2013 shared task (Nédellec et al., 2013) data. The BioNLP data is described in more detail in chapter 10. In the first task (GE task) of BioNLP, an anaphora resolution task was integrated. The annotations given in the data sets were extracted and a pure anaphora resolution corpus was created from them. The anaphora resolution system was evaluated on the development data set (the gold

standard of the test data set was not yet released). This data set consisted of 826 sentences from 9 different publications.

After a first evaluation of the system, however, it became obvious that the BioNLP annotation was incomplete. Not all anaphoras occurring in the sentences were properly annotated. This lead to cases where our algorithm delivered correct results that were missed in the annotation and thus would be counted as false positives. Take for example the following passage:

> We cloned a 1025 bp promoter, which ranges from position - 959/+66 relative to the identified TSS. Luciferase reporter assays showed that this promoter was transcriptionally active in A3.01 T cells.

The constraint-based anaphora resolution algorithm correctly resolved 'which' and 'this promoter' to 'a 1025 bp promoter'. In the BioNLP corpus, however, there were no anaphoras annotated. Including such cases would have distorted and worsened the evaluation results. For this reason, sentences containing anaphoras that were not annotated were removed from the evaluation data set. In addition to that in some cases BioNLP annotated anaphoras that were not actual ones. In these cases, usually a class term immediately preceded a biological entity like in 'the cytokines interleukin (IL)-4, IL-5, and IL-13'. Here, the noun 'cytokines' serves as an explanatory attribute of the following terms. In some cases, BioNLP, however, annotated such formulations as anaphoras. These instances were removed as well. Both of these filtering procedures resulted in a reduced final set of 687 sentence containing 95 anaphoras.

Table 6.1: Performance of anaphora resolution system on BioNLP data.

Task	Accuracy	Precision	Recall	F-measure
Trigger detection		0.74	0.37	0.49
Antecedent resolution	0.53			
Anaphora resolution		0.39	0.18	0.24

For anaphora resolution as well as its subtasks trigger detection and antecedent resolution performance values were calculated. For antecedent resolution, the accuracy was calculated instead of precision, recall and F-measure values since triggers were only detected when an antecedent was found. Thus, the missed antecedents in cases of found triggers are already incorporated in the trigger detection values. The results can be seen in Table 6.1.

As can be seen the trigger detection is fairly reliable with a precision of 0.74. Examining the remaining errors showed that all of them were due to utterances describing an individual but containing a trigger word like 'the p53 protein'. These cases can easily be resolved by constraining the trigger detection process. However, a reliable ontology is required in order to distinguish such cases from other modifiers, that can occur before trigger words (e.g. like in 'the activated protein' or 'the virus protein').

The recall values are lower than the precision values as would be expected in a rule-based system. More than half of the trigger words were resolved correctly. Thus, if the trigger detection is resolved as mentioned above the anaphora resolution system contributes to increasing the performance of the text mining system it is used in. Yet, the overall performance values are not completely satisfying. In order

to detect possible starting points for improvements of the system, an error analysis with special focus on false negatives was conducted. This error analysis revealed the following error sources:

- As mentioned before, the syntactic role constraint was planned but not yet included in the system due to the lack of information in the topic maps. The error analysis showed that including this constraint would prove beneficial. For example, in the following passage the system falsely resolved 'its' to 'transcription factor' instead of 'FOXP3' which would not have happened if the syntactic role constraint would have already been used.

  > FOXP3 is an essential transcription factor for natural, thymus-derived (nTreg) and inducible Treg (iTreg) commitment; however, the mechanisms regulating its expression are as yet unknown.

- The anaphora resolution system increases its performance with the quality and especially the extent of the used ontology and event extraction system. The more terms one can categorize within the ontology, the better the constraints of the anaphora resolution system take effect. The lack of a very large, comprehensive ontology consequently caused errors.

- The Clueda sentence analysis framework does not yet resolve all kinds of appositions properly. Since there are often anaphors or antecedents in appositions, this is another source of error. An example of this can be seen in the following utterance:

  > ... two PKD isoforms, PKD1 and PKD3, ...

  Here, 'PKD1 and PKD3' is the apposition that contains two antecedents the anaphor 'two PKD isoforms' should be resolved to. The lack of syntactic information in the topic maps, however, avoids a proper resolution of this anaphora.

- Some peculiarities of the annotation additionally caused a decrease in the observed performance values. Exemplarily, in the following sentence 'interferons' was considered a trigger word and resolved to two before mentioned interferons. Depending on how one interprets this statement, however, it either means that not further specified interferons caused the upregulation or that all interferons did so. Yet, in both cases 'interferons' does not reference specifically the two before mentioned ones, which would be the case if the statement included 'these interferons' instead of the version without a determiner.

  > It has been described previously that A3G gene expression is upregulated by interferons in hepatocytes and macrophages (46-48 , 52) .

  In other cases, the anaphora was not fully resolved but instead another anphor was annotated to be the antecedent. An example of this is given in the following passage:

  > Thus, the induction we observed was most likely mediated by the described interferon-responsive elements. However, according to our results, these motifs can enhance transcription in hepatic cells, but not in T cells.

  Here, the BioNLP annotation resolves 'these motifs' to 'the described interferon-responsive elements' instead of the actually described interferon-responsive elements mentioned earlier in the

text. This annotation procedure differs from the strategy followed in the system described in this chapter. Furthermore, it seems pointless to resolve anaphors to other underspecified terms considering that the aim of anaphora resolution is to resolve such underspecified terms to the actual entities they reference. A proper anaphora annotation would recognize this chain correctly. These and other peculiarities (e.g. the approach to metonyms) reduced the meaningfulness of the evaluation results and decreased the obtained values.

- Very complex structures need more sophisticated constraints. For example, the term 'family member' includes two anaphoras at once. First the family that is meant needs to be determined. Based on this the ontology constraints of the member can be deduced, which then needs to be resolved in a second resolution step. Utterances like this accounted for additional errors of the system.
- The system lacks an algorithm to distinguish between expletive and normal pronouns. Expletive pronouns (also called dummy pronouns) are pronouns that have an exclusively syntactical function but do not contribute to the semantics of a sentence and thus should not be resolved to an antecedent. Expletive pronouns occur quite frequently in scientific texts in expressions like the following:

  > It has been shown ...

  Here, the 'It' is the expletive pronoun, which produces a false positive if it is resolved. Since for the word 'it' the expletive pronoun seemed to occur more frequently than the normal one, it was decided to leave 'it' out of the vocabulary of pronouns that were resolved. Consequently, in the cases where normal forms of 'it' were used, false negatives occurred. This could be corrected in future work by building a simple expletive pronoun detector that checks for the verbs the pronouns are occurring with.
- Finally, as with every high-level linguistic algorithm, the quality of the underlying natural language processing tools influences the quality of the system. Errors in sentence splitting, tokenization, POS tagging, chunking, PAS extraction and event detection diminished the quality of the anaphora resolution system. Since the anaphora resolution system described here is based on a prototype such errors still occurred comparatively frequently. Once the surrounding system is more mature and correcting procedures like the backtracking described in chaper 10 are included, the quality of the anaphora resolution is likely to increase accordingly.

6.4 Conclusion

Summing up one can state that based on the obtained results the anaphora resolution system still requires additional modifications in order to reach the full potential of the approach. The shortcomings that caused these results, however, could be explained by the lack of required information in the input topic maps, the too sparse ontology and event extraction system and a required distinction algorithm between expletive and other pronouns. Each of these could be corrected with additional future work leading to the conclusion that the approach in general seems promising. Cases in which these missing functionalities and information were not required to solve the anaphora were already resolved with

reasonable precision. The situation that an algorithm automatically improves its performance when additional functionalities and information are added to the system corresponds to the supersemantic paradigms promoted in this thesis.

6.5 Related Work

An overview over related related anaphora resolution approaches is given in Table 6.2

Table 6.2: Text contextualization: Related work

Authors	Year	Approach	Domain
McCarthy and Lehnert (1995)	1995	Decision trees	Business
Soon et al. (2001)	2001	Decision trees	Business
Ng and Gardent (2002)	2002	Decision trees	Business
Uryupina (2010)	2010	Support vector machine	Multilingual
Culotta et al. (2007)	2007	First-order probabilistic model	Various
Morton (2000)	2000	Maximum-entropy classifier	Politics, Business
Bengtson and Roth (2008)	2008	Perceptron	Politics
McCallum and Wellner (2005)	2005	Conditional random fields	Politics, Business
Ng (2008)	2008	EM clustering	Politics
D'Souza and Ng (2012)	2012	Rule-based	Biomedical
Gasperin and Briscoe (2008)	2008	Bayes classifier	Biomedical
Su et al. (2008)	2008	Augmented learning	Biomedical

7

Corpus Contextualization

A corpus is commonly the largest unit of textual information. It can incorporate arbitrarily vast collections of texts and hence is not limited in its size. Because of this corpora are especially interesting for assessing the statistics of underlying units, like words or relations. In this chapter, a brief overview over the information that can be extracted from a corpus is given. Furthermore, two methods to use such information and their application to support building a dictionary and to visualize large amounts of text mining results respectively are presented.

7.1 Corpus Information

The linguistic field that typically deals with corpus information is corpus linguistics. Corpus linguistics is often distinguished from classical linguistic approaches, since it is more practically oriented. Instead of focussing on the theoretical framework of language, it investigates how language is practically used (Gries, 2010). Thus, its methodology is based on analyzing large collections of real texts, speeches or conversations (Dash, 2010). Or as McEnery and Wilson provocatively put it: "Corpus Linguists study real language, other linguists just sit at their coffee table and think of wild and impossible sentences" (McEnery and Wilson, 2005) [7]. These corpora are analyzed with computerized empirical methods to determine quantitative features of the given collection. The features, in turn, are then qualitatively interpreted to determine their function (Biber et al., 1998).

Because of the large amount of textual data statistical evaluations are possible. These were initially used to count frequencies of words, word combinations or parts of speech. More elaborate studies extended this to detect more complex patterns within the corpus. Such patterns are called association patterns and they describe how different linguistic entities, like words or grammatical structures,

[7] It should be mentioned that Enery and Wilson renounce this view as folklore.

are associated with each other or with non-linguistic features within the given corpus. These non-linguistic features include the distribution of patterns within different text collections, e.g. the different distributions within review and research articles or within papers from different domains. Futhermore, the distinction between text collections in corpus linguistics has often been made based on dialect, time period, and register (alternations of language motivated by different social situations, e.g. very casual expression while talking to friends) (Biber et al., 1998). A possible use of such association patterns is the bootstrapping method that was already described in chapter 5.

The early approaches to biomedical text mining were strongly influenced by corpus linguistics. Especially, the co-occurrence analysis - or collocation how it is often referred to in corpus linguistics - of terms was and in many cases still is applied to associate biological entities. In such an analysis terms are associated with each other when they co-occur frequently within a certain window of text. One can try to infer the type of association by looking at co-occurring terms describing such types, e.g. 'activates' or 'inhibits' in the biomedical context. Such approaches were used to collect meaningful biological associations, like protein-protein interactions.

Most of the association patterns detected nowadays contain implicit information. They are rather vague and do only allow a statistical interpretation but not a precise, logical one. With text mining methods evolving, however, more and more explicit and reliable information about real biological events and their contexts will be extracted. Combining this information on a corpus level requires adequate resoning capabilities. While applications in this direction are still scarce (see (Tari et al., 2010) for one of the few current applications), it has alredy been pointed to (e.g. in (Ananiadou et al., 2010)) that the future of text mining might be an integration with automated reasoning.

Automated reasoning is part of the field of artificial intelligence and is deeply rooted in logics. It originated from the work of Newell, Shaw and Simon who first created a program called the Logic Theorist. This program was able to formally prove thrity-eight theorems presented in the Principia Mathematic by Rusell and Whitehead, including in one case a more elegant proof than the original one (Lewis, 2000). Automated reasoning methods are currently mostly used for mathematical proofs, error checking of software code and communication protocols as well as the synthesis of knew knowledge by reasoning in knowledge bases and ontologies (Konev et al., 2010). In the context of biomedical text mining, automated reasoning might be used among others in the generation of new hypotheses and the detection of conflicting information as well as logical fallacies in the argumentation of scientists.

Beside the analysis of corpora, the construction of them is the second important task in corpus linguistics. This process subsumes the compilation of the text collections but often more importantly also the annotation of them. Corpora are often annotated with part of speech tags. On top of this, syntactic or semantic structures are frequently annotated. In this case the corpora are commonly referred to as treebanks due to the tree structure of syntactic and semantic analyses. Among the most well known corpora are the Brown corpus (Francis and Kucera, 1979) and the Wall Street Journal corpus (Paul and Baker, 1992). An example of a syntactically annotated treebank is the Penn Treebank (Taylor et al., 2003) that is among others based on both the Brown and the Wall Street Journal corpus. An example of a semantically annotated treebank is the Propbank (Kingsbury and Palmer, 2002) which is in turn based on the Wall Street Journal section of the Penn Treebank.

Annotated corpora are a necessary requirement for a variety of machine learning based NLP tools. Sentence detectors, POS taggers as well as syntactic and semantic analyzers all take their training

examples from usually manually annotated corpora. Exemplarily, the Senna tool that is the foundation of Excerbt is trained on the Propbank corpus.

7.2 N-gram Analysis

Creating ontologies and dictionaries is a very time-consuming task. It is necessary, however, to have a knowledge base, from which one can identify named entities, in order to perform most text mining analyses. Hence, approaches that quicken the ontology creation process can be very useful. One of these approaches coming from corpus linguistics is N-gram analysis. It is frequently used in ontology learning (see e.g. (Hazman et al., 2011)) and has also been used in the context of Excerbt by Robert Strache in his Master's Thesis (Strache, 2012).

N-grams are collections of N consecutive words that frequently occur together in a corpus. The idea behind N-gram analysis for ontology learning is to look for the most frequent N-grams consisting of certain parts-of-speech. Expecting the most frequent ones to also be the most important ones, one can then use these directly or after manual verification as dictionary or in the hierarchical ordering of an ontology.

There are many variations on how to exactly perform this analysis. The one used together with Excerbt so far is exemplarily described in the following. Here, only unigrams to 4-grams consisting of adjectives, nouns, gerund forms, determiners and prepositions were used. Unigrams were restricted to nouns, while the others needed to start or stop either with a noun, adjective or gerund form. Furthermore, to increase the frequencies stemming was employed and to avoid non-specific terms stop words were removed (Strache, 2012).

In the context of the rare diseases analysis (see chapter 5) a similar but modified version of this algorithm was used. Here, all word sequences starting and ending with a noun consisting of nouns,adjectives, the genetive 's', the genetive 'of', and the determiner 'the' were extracted. This way sequences of arbitrary length N could be extracted. For the rare diseases analysis, every N-gram that occurred at least seven times was manually checked whether it described a symptom.

7.3 Word Space Models

While N-gram analysis might answer the question of which terms are important in a corpus, it cannot say anything about their meaning. A corpus linguistic approach to term meaning are word spaces. In a word space, meaning is modelled as a multi-dimensional space where words with similar meanings are close to each other. Thus, every word is modelled as a multi-dimensional vector that represents its meaning. The vectors are derived by considering in which contexts in the corpus they are occurring. Or as John Rupert Firth put it: "You shall know a word by the company it keeps" (Firth, 1957).

The word-space model was introduced by Magnus Sahlgren in his doctor thesis (Sahlgren, 2006). The foundation of the model is the so-called distributional hypothesis, which states:

> "Words with similar distributional properties have similar meanings."(Sahlgren, 2006)

The hypothesis has been experimentally validated several time (McDonald and Ramscar, 2001; Miller and Charles, 1991; Rubenstein and Goodenough, 1965). For instance, Rubenstein and Goodenough (Rubenstein and Goodenough, 1965) found a correlation between distributional features of terms and synonym judgements of university students. There exist different methods to create word space models. Among the most popular ones are latent semantic indexing and random indexing.

Latent semantic indexing (LSI) , or latent semantic analysis (LSA) , was introduced by Deerwester et la. in 1990 (Deerwester et al., 1990). In this method the occurrence frequencies of terms with respect to different documents of a large corpus are counted. The result of this is a very large and often very sparse term-document matrix. This matrix is made up of vectors, each representing the distributional features of a term within the corpus. Additionally, the frequency values are weighted according to the importance of the respective words in their context. However, since it is very impractical to work with such a large matrix, it is decomposed into its so-called principal components. These principal components are more concise representations of the term vectors that should still capture the important distributional features. In latent semantic analysis, this dimensionality reduction is often performed by a linear procedure called singular value decomposition (SVD) (Landauer and Dumais, 1997; Landauer et al., 1998).

SVD was first introduced in latent semantic analysis. It decomposes the original matrix into three individual ones, where the product of the three matrices is the original one. One of the matrices contains the information of the rows of the original matrix in the form of orthogonal factor values. Another one does the same for the columns. The middle matrix is a diagonal matrix responsible for scaling the factors correctly in order to obtain the original matrix when multiplying. Applying this analysis results in a matrix describing terms and another one describing documents. This is why latent semantic analysis can be used to create word spaces as well as document spaces. Since the scaling of the factors can be seen in the diagonal matrix, the dimensionality reduction can be accomplished simply by deleting the smallest factors in it. This reduction results in a least-squares best fit of the original matrix (Landauer et al., 1998). The resulting reduced word vectors then make up the word space.

One drawback of SVD is that it is computationally expensive and since it is its very purpose to be applied to very large corpora, this can become an issue in many applications. For this reason, approximate alternatives to LSA were constructed that avoid this problem. One of these methods is random indexing (RI). RI omits the time-consuming decomposition by reversing the order of the collection of distributional features and the dimensionality reduction. In RI, the first step is to randomly create sparse document vectors that already have the desired dimensionality. These vectors, called index vectors, are unique and contain only values of 0, +1 and -1. Then the term vectors are created by summing up the document vectors in which the term occurs. RI is only an approximation of LSA, since the randomly chosen vectors are usually not orthogonal. Based on an observation of Hecht-Nielsen (R., 1994), however, it is possible to approximate orthogonality in this way because there are so many nearly orthogonal directions in a very high-dimensional space (Sahlgren, 2005).

7.4 Word Space Visualization of Text Mining Results

Text mining of large collections of texts often returns large collections of results. These results when represented in the form of a graph produce a so-called "hairy ball". This term refers to a highly-connected graph that is too large to be intuitively interpretable by humans. Thus more time or more detailed analyses are required to understand the text mining results properly. In many situations, however, it would be valuable to get a quicker grasp of the returned results. For instance, if one wants to get a quick overview or if one is new to a field, a more intuitive representation of text mining results would be useful.

Based on these considerations I designed and developed a visualization system utilizing word spaces (a paper describing the results is currently in revision for publication in the Journal of Biomedical Semantics (Blohm and Meiners, 2014)). The underlying idea is to focus on one or a few terms and display all relations found by a text mining system. In order to ensure the clarity of the displayed information, the related concepts are clustered and only the most important ones are shown. This approach is based on a human's natural reaction to complexity. First the information is ordered, then it is prioritized. The ordering process uses the word space model. The prioritization process is based on frequency counts, which as mentioned are frequently employed in corpus linguistic applications.

The result of this visualization is a graph like the one seen in Figure 7.1. At the center of the graph is the search term, for which all relations are shown, in this case the 'proteasome'. The entities with which this search term interacts are clustered in semantic groups. The clustering is performed on the basis of the Semantic Vectors Package (Widdows and Ferraro, 2008) implementation of LSI and RI. The derived semantic vectors are then clustered using a k-means clustering algorithm from the same package. One can see that semantically related terms occur in the same clusters. For instance, 'MHC' and 'Antigens' are clustered together. The major histocompatibility complex (MHC) molecules have the function to present antigens to T cells in order to indicate whether foreign proteins have entered the cell.

The amount of terms in each cluster is indicated by the size of the node in the graph. Additionally, the bigger clusters are colored in green, the smaller ones in purplish blue. The displayed terms of each cluster correspond to the ones that have the most connections to the central search term. Thus, the prioritization here is conducted by frequency counts. Analogous to the nodes the edges give a hint about the frequency properties of the underlying data. The more evidences (relations found by the text mining system) point to a connection between the search term and the cluster, the larger is the edge between them. Furthermore, stronger connections are displayed in green, more moderate ones in yellow and weak ones in red.

This presentation should give the user an intuitive overview about the search term and its relations to other biological entities. For instance, for the term proteasome the strongest connection is to the cluster in which proteins and degradation are prioritized. This makes sense since the main function of the proteasome is the degradation of proteins. To investigate the results further, the user can then click on one of the clusters. By doing so the terms belonging to this cluster are presented in the same form. If the amount of them is large, they are again clustered and prioritized in the same manner. If the amount is small, each term is represented by its own node. Additionally, information about the sources of the relations is available to the user of the tool. By clicking on the edges on any level of the graph all evidences for this edge are displayed. An evidence consists of the sentence from which the

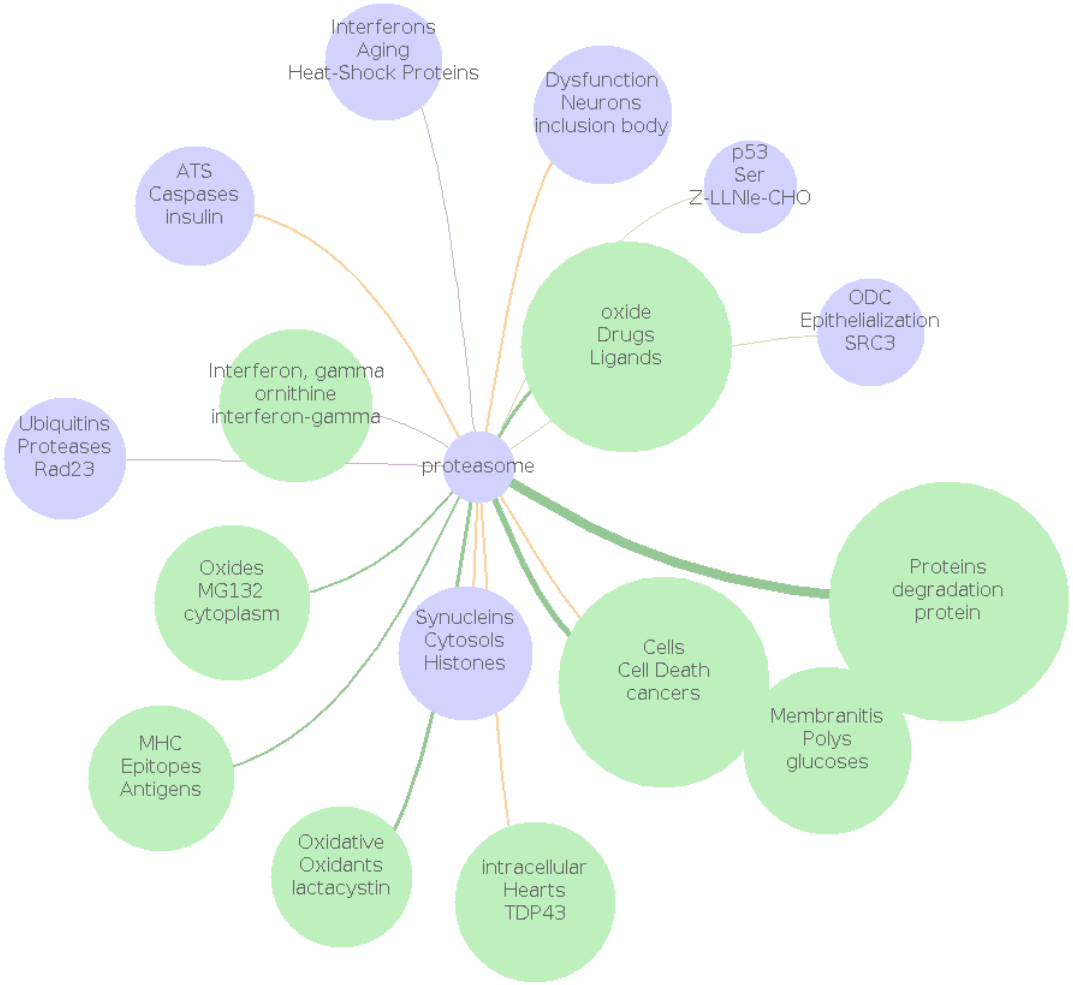


Figure 7.1: Visualization of Excerbt text mining results of the term 'proteasome'.

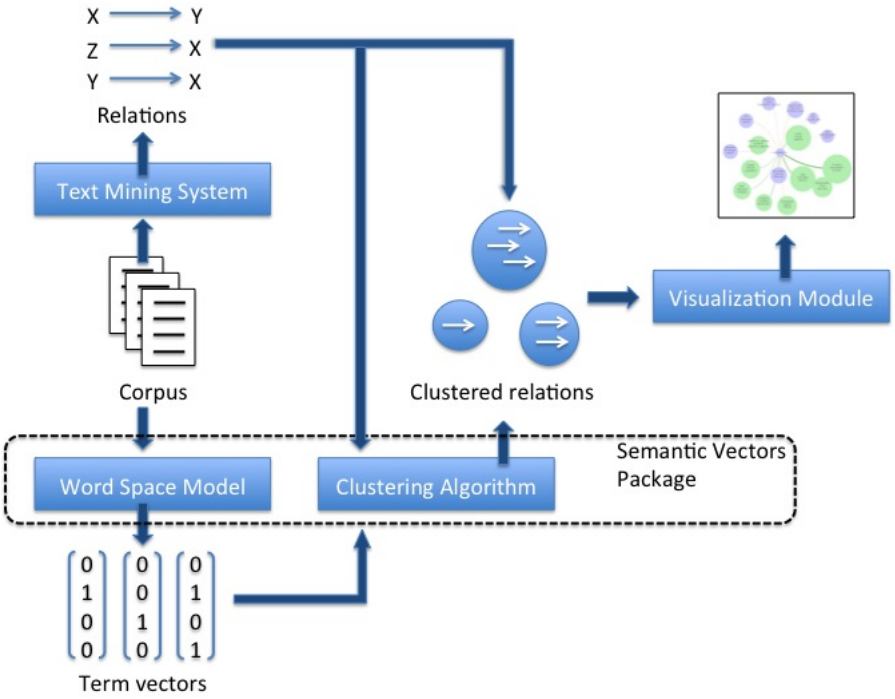


Figure 7.2: Overview of the architecture of the visualization tool.

text mining system extracted the relation and a link to the corresponding PubMed or PMC publication. Using this information the user can check for errors of the text mining system or gather additional context information about the relation.

Figure 7.2 illustrates how the different components of the system interact. As can be seen, the analyzed corpus is the origin of the analysis. Both the word space model and the text mined relations are extracted from it. The corpora for the two does not need to be the same. It makes sense, however, to choose corpora from the same domain in order to properly represent the domain specific aspects of language. The terms in the text mined relations are then clustered according to their term vectors resulting in clustered relations.

The system was evaluated using text mining results from Excerbt for the terms 'proteasome', 'COPD' and 'IPF' . However, in order to increase the applicability of the tool binary relations from any text mining tool can be visualized with it. For this purpose, a generic input format was defined. Finally, the clustered relations are presented to the user through the visualization module. Each of the clusters is labeled with the most frequent terms for prioritization. For visualization the Prefuse (Heer, 2005) Java-library was used.

The tool is designed to explore text mining results for a certain topic. The visualization is centered around this topic. This should be especially useful, when one wants to become acquainted with a certain field or just wants to get a quick overview. A second application scenario is comparing multiple concepts. Here, the direct and indirect (over one other node) connections of two concepts are important. To allow the visualization tool to be useful in these situations a second visualization mode

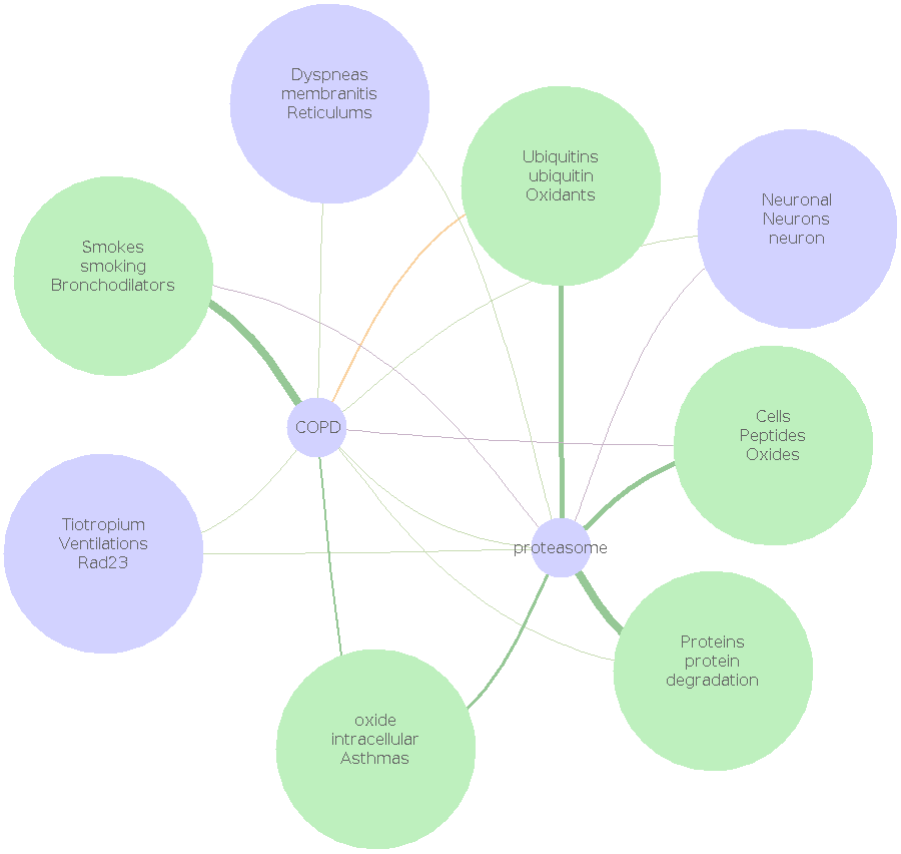


Figure 7.3: Multi-concept mode of the visualization tool.

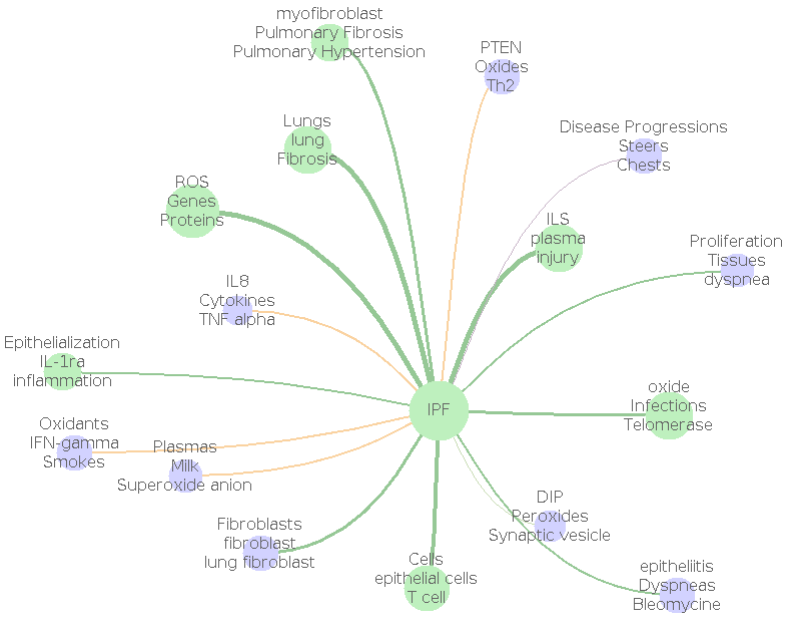


Figure 7.4: Visualization of the text mined results for IPF.

was implemented. In this mode, there are several topics at the center of the visualization. The rest of the implementation stays the same. An example of such a visualization is given in Figure 7.3.

7.5 Evaluation

Evaluating a visualization tool is always difficult since there are rarely any easily accessible objective measures. The optimal solution to this usually is an extensive usability study. However, for this a large amount of test persons is required, which often is not available to the developer of the tool. Likewise, in this work we lacked the resources for such an extensive evaluation. Alternatively, different indicators were considered, which should give a comprehensive picture of the way the tool is working.

First the ability of the tool to order the relations in a meaningful way was explored. For this purpose the results of the semantic clustering were investigated in more detail. One peculiarity that catches the eye just by browsing through the results is the clustering of synonyms. The synonym module of Excerbt was used to consolidate synonyms. The tool, however, showed that the module misses a lot of synonyms. As can be seen in Figure 7.4, synonymous terms or more general terms like ’lung’ and ’Lungs’ or ’Fibroblasts’, ’fibroblast’ and ’lung fibroblast’ are not mapped onto each other. Interestingly, however, the terms are very often clustered together. This opens up the possibility of an additional application field for the presented tool. It might be useful when working with different vocabularies to support the mapping of identical but differently named concepts. Furthermore, the clustering of synonyms is an indicator that the semantic clustering is working in the intended way. Synonyms are semantically identical concepts. Thus, in a semantic clustering approach, they are supposed to be clustered together.

A classical way of ordering concepts in a semantic manner is the use of an ontology. Ontologies are typically hierarchically organized and the entities within them are standing in a is_a or part_of relation. In many cases, however, ontologies are not available or do not cover a desirable range of relations. Therefore the presented visualization tool offers an alternative to the use of established ontologies. It considers all relations found by a text mining system and clusters them according to distributional features. Since both ways of orderings are semantic, however, there should be a correlation between both. In the progress of this work, such a possible correlation was investigated. The terms' memberships in different clusters or ontology classes are nominal features, because the different clusters and ontology classes (as long as one considers classes of the same level) are not in a particular order. A typical way to measure associations in nominal data is an association measure called Cramer's V.

Cramer's V is based on the work of the swedish statistician Harald Cramer. It approximates how much of the values of a group of nominal variables can be ascribed to the association of the second group of nominal variables. In this cases it states how much of the semantic clustering based on the word space model of the visualization tool can be explained by its association with a given ontology. The measure was calculated for the two different approaches to creating word space models. The Excerbt ontology was used for the association. This ontology is composed of a variety of other ontologies like MeSH terms, Entrez Gene and many others. The results of the association analysis can be seen in Table 7.1.

Table 7.1: Association scores for different word space approaches.

Association clustering - ontology		
Graph	Avg. Cramer's V Random Indexing	Avg. Cramer's V Latent Semantic Indexing
Proteasome	0.444	0.462
COPD	0.462	0.494
IPF	0.364	0.389
Avg.	0.423	0.448

The association scores are between 0.364 and 0.494. Thus, between a third and half of the semantic clustering can be explained by the association with the ontology. Furthermore, the LSI values are slightly higher than the ones of RI. This was to be expected, since RI is only an approximation of LSI. All in all the association analysis confirmed that the semantic clustering works appropriately. A substantial portion of the ordering in the Excerbt ontology can be seen in the clusters derived with this different method. However, the semantic clustering also provides information not contained in the Excerbt ontology and thus can be seen as a valuable complement.

Both considered indicators support the proper functioning of the semantic clustering. Additionally, the prioritization process was investigated in more detail. The goal of the prioritization is that the most important information about the investigated topic is displayed on the highest level. As a ground truth for what the most important information about a topic might be we considered official definitions of terms. If the prioritization approach works as intended, it should prioritize the terms used in these definitions. We applied this evaluation using the MeSH definitions of the terms 'proteasome', 'COPD' and 'IPF'. MeSH defines these terms as follows:

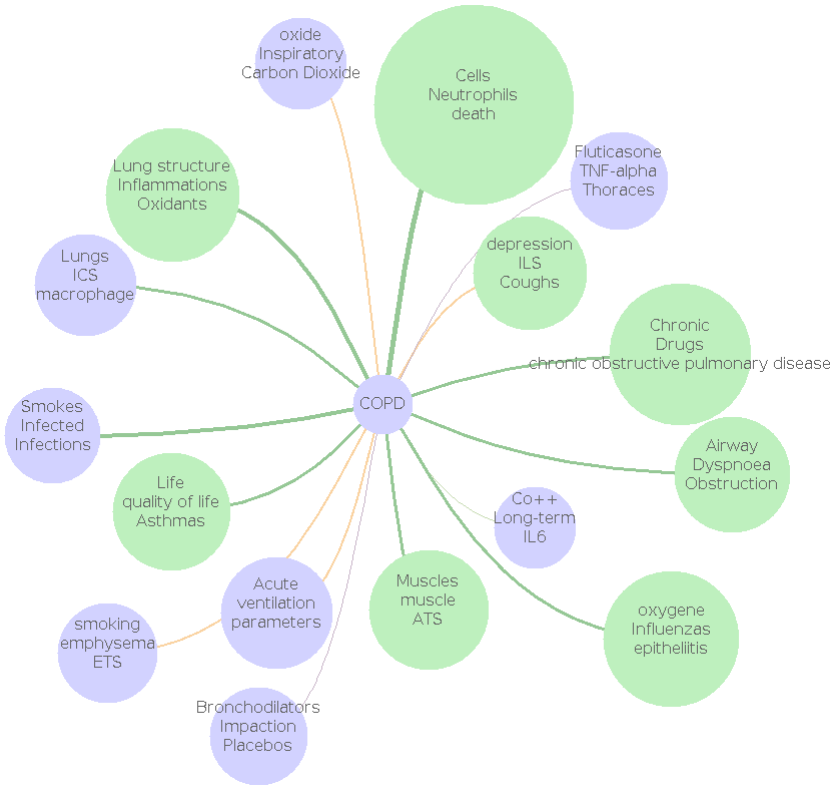


Figure 7.5: Visualization of the text mined results for COPD.

"Proteasome Endopeptidase Complex: A large multisubunit complex that plays an important role in the **degradation** of most of the cytosolic and nuclear **proteins** in eukaryotic cells. It contains a 700-kDa catalytic sub-complex and two 700-kDa regulatory sub-complexes. The complex digests **ubiquitinated proteins** and **protein** activated via ornithine decarboxylase antizyme" (http://www.ncbi.nlm.nih.gov/mesh/68046988).

"Pulmonary Disease, Chronic Obstructive: A disease of chronic diffuse irreversible **airflow obstruction**. Subcategories of COPD include CHRONIC BRONCHITIS and PULMONARY EMPHYSEMA" (http://www.ncbi.nlm.nih.gov/mesh/68029424).

"Idiopathic Pulmonary Fibrosis: A common interstitial **lung** disease of unknown etiology, usually occurring between 50-70 years of age. Clinically, it is characterized by an insidious onset of breathlessness with exertion and a nonproductive cough, leading to progressive **DYSPNEA**. Pathological features show scant interstitial **inflammation**, patchy collagen **fibrosis**, prominent **fibroblast proliferation** foci, and microscopic honeycomb change" (http://www.ncbi.nlm.nih.gov/mesh/68054990).

The terms or their synonyms that were prioritized the way that they appeared on the highest level of the visualization (see Figures 7.1, 7.4 and 7.5) are shown in bold. As one can see, for all three concepts important terms of the definition appear on the highest level of the visualized graph. If one looks more closely, one can even observe what kind of terms appear. The visualization tool seems to

cover the functions of the terms or, in case of the diseases, the resulting phenotypes very well: for the proteasome its main function, the degradation of proteins, is found; for COPD the main symptome of airflow obstruction; and for IPF even multiple phenotypes.

On the other hand, two types of information seem to be left out quite consistently. First, the direct definition is missing: neither for the proteasome the large multisubunit complex shows up nor for the other two the term disease (although disease progression appears for IPF). Secondly, the parts of which the term is made up of are not accounted for. In case of the proteasome the sub-complexes are not mentioned and in case of COPD the subcategories. These shortcomings can, however, be explained quite easily, when considering that relations taken from Excerbt were used. Definitions are commonly expressed in is_a relations and sub-system relationships in the form of part_of relations. Excerbt, however, does not cover these two relations. Consequently, it had to be expected that they would not show up in the visualization. Using a more comprehensive text mining system instead of Excerbt would probably circumvent this shortcoming. In order to make this possible the tool was implemented the way that relations given in a generic format coming from any text mining system could be used as input.

7.6 Conclusion

The practical approach of corpus linguistics is well suited for the biomedical text mining. The association patterns that can be found in corpora can quicken the necessary construction of ontologies and event definitions. Furthermore, the quantitative analysis of them can be used to order concepts and relations. In this chapter, two practical applications of corpus linguistic approaches were presented. The N-gram analysis proved useful in establishing a vocabulary, while the word space model could help consolidate vocabularies by detecting synonyms. Furthermore, the word space model was used as the basis for a visualization tool that helped represent text mining results in a more comprehensible way.

Future developments of biomedical corpus linguistics might lead to a use of methods from the field of automated reasoning. A more comprehensive discussion on how such a development might look like is given in the discussion. Furthermore, the association patterns might become more complex and better fit the needs of biomedical researchers as biomedical text mining evolves.

7.7 Related Work

An overview over related related corpus contextualization approaches is given in table 7.2

Table 7.2: Corpus contextualization: Related work

Authors	Year	Approach	Domain
Felizardo et al. (2010)	2010	Similarity-based corpus visualization	Software engineering
Fortuna et al. (2005)	2005	LSI-based visualization	Research projects
Malheiros et al. (2007)	2007	Visual Text Mining	Software eng.
Nikitin et al. (2003)	2003	Pathway visualization	Biomedical
Kemper et al. (2010)	2010	Pathway visualization	Biomedical
Hazman et al. (2011)	2011	N-gram ontology learning	Agriculture
Sanchez and Moreno (2004)	2004	Bi-gram ontology learning	Biosensors
Cimiano et al. (2005)	2005	Formal Concept Analysis	Tourism and finance
Dezhkam and Khalili (2013)	2013	Pattern-based ontology learning	Biomedical

8

Integration of External Knowledge

Text mining focuses on unstructured or semi-structured information. In order to get a full picture of the existing knowledge, however, it has to be integrated with all the information. This includes the structured information from databases. In the biomedical domain, this first and foremost experimental results are stored in structured resources and often only partially described in publications. The missing data can either be taken directly from experiments or from structured resources, like IntAct (Kerrien et al., 2012) and others, that collect experimentation results. A good example of such an integration is String (Franceschini et al., 2013), which among others integrates text mining results and results from coexpression and high throughput experiments.

The integration of text-mined and structured data with the purpose of reaching a better understanding is a supersematnic task. The reasoning on the basis of both data sources can either happen automatically or by a human. In this chapter a tool combining structured and text-mined information for an improved human experiment interpretation is introduced. The tool focusses on gene enrichment analysis. It facilitates the interpretation of the results by providing additional information from different text mining resources. It was implemented under my supervision in the course of the Bachelor Thesis of Tim Jeske (Jeske, 2013). I designed the system and supervised the development process as well as the GO analysis and the practical application.

8.1 Functional Analysis of Gene Lists

The introduction of high-throughput experiments fundamentally changed the way genes were analyzed. Instead of focussing on single genes, analyses on the whole genome became possible. This immensely increased the range of single experiments and lead to a less biased approach, since the considered genes were no longer predetermined by the researchers opinions and intuitions. On the other hand, the high-

troughput approach increased the challenges that were presented to the respective researcher. Before he was able to acquire deep expert knowledge of the limited domain he was investigating. Afterwards, the interpretation of his high-throughput results required him to have far-reaching expertise about all the very different genes that might prove significant in the experiment.

With the use of microarrays many thousands of genes could be analyzed at once. This made it impossible for the scientist interpreting the experiment to be an expert on each one of them. In order to nevertheless facilitate an appropriate interpretation different resources were created to support the scientist. These resources order the genes according to different criteria like which known pathways they belong to, which properties they have or which functions they fulfill. The most prominent resource for pathways is the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). The most prominent functional ordering of genes is given by the Gene Ontology (GO) (Ashburner, 2000).

The Gene Ontology consists of three subparts: the biological process ontology, the molecular function ontology and the cellular compartment ontology. GO defines the three categories as follows:

> *"Biological process refers to a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. Processes often involve a chemical or physical transformation, in the sense that something goes into a process and something different comes out of it. Examples of broad (high level) biological process terms are 'cell growth and maintenance' or 'signal transduction'. Examples of more specific (lower level) process terms are 'translation', 'pyrimidine metabolism' or 'cAMP biosynthesis'."* (Ashburner, 2000)

> *"Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. This definition also applies to the capability that a gene product (or gene product complex) carries as a potential. It describes only what is done without specifying where or when the event actually occurs. Examples of broad functional terms are 'enzyme', 'transporter' or 'ligand'. Examples of narrower functional terms are 'adenylate cyclase' or 'Toll receptor ligand'."* (Ashburner, 2000)

> *"Cellular component refers to the place in the cell where a gene product is active. These terms reflect our understanding of eukaryotic cell structure. As is true for the other ontologies, not all terms are applicable to all organisms; the set of terms is meant to be inclusive. Cellular component includes such terms as 'ribo-some' or 'proteasome', specifying where multiple gene products would be found. It also includes terms such as 'nuclear membrane' or 'Golgi apparatus'."* (Ashburner, 2000)

The terms within the different ontologies can be in different kinds of directed relations to each other. As common in ontologies there exist is_a and part_of relations. Furthermore, GO comprises regulation relations. These can be further specified as positive or negative. Apart from that a NOT qualifier exists that indicates that a function, which might be intuitive, was tested not to exist. Finally, each entry in GO is accompanied by an evidence code. Such a code denotes how the entry was created. Here, it is distinguished whether this was automatic or with the contribution of manual curators and whether it was based on an experiment, computational analyses or merely statements.

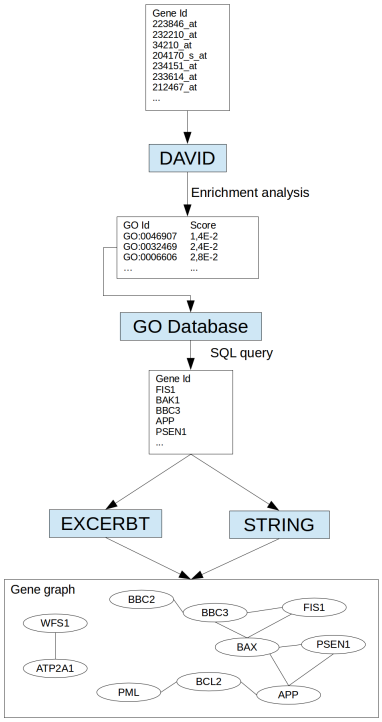


Figure 8.1: Overview of the workflow of the text mining assisted functional analysis tool. Picture taken from (Jeske, 2013).

Functional orderings like KEGG and GO can be used to facilitate the interpretation of high-throughput experiments. The results of such experiments are usually a list of significant genes. A common strategy then is to associate these genes with functions by looking at the categories in which these genes fall. The categories that contain a significant amount of genes from the input list are called enriched and considered relevant for the interpretation of the experiment. This process is automated in different tools. In this work, the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al., 2003) was used.

DAVID comprises a knowledge base of different resources and a variety of tools. For the further analysis the gene annotation enrichment analysis was used. This analysis associates gene lists with GO terms by using a modified Fisher's exact test. The communication with the tool was implemented using the web services of DAVID (Jiao et al., 2012).

Tools like DAVID can give the scientist a general idea of the relevant functions. If the interpretation leaves it at that, however, it remains rather superficial. Thus, in order to allow a more profound interpretation of the experimental results again expert knowledge of the scientist is needed. Consequently, the quality of the analysis is based on how detailled the knowledge of the corresponding researcher is. This can again lead to rather shallow interpretations or to researchers only focussing on those areas which they are familiar with. The latter again introduces a bias towards the background knowledge of the person interpreting the experiment. In order to further enable scientists to perform a

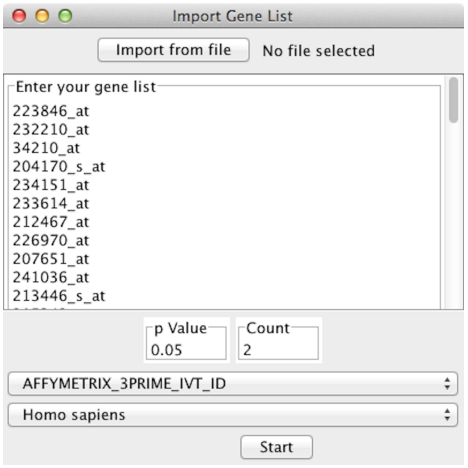


Figure 8.2: Input window of the tool. Picture taken from (Jeske, 2013).

more comprehensive and unbiased analysis more detailled information about how the genes that are subsumed in the gene sets actually interact with each other would be required. For this purpose, in this work a tool offering a more detailled description of the relevant GO categories on the basis of text mining is introduced.

8.2 Functional Analysis Using Text Mining

The idea behind the functional analysis tool developed in this work is to provide more context information to the user by integrating text mining results into a gene annotation enrichment analysis. For this purpose several existing tools and resources were integrated into one comprehensive system that should facilitate a more informed interpretation. In order to maximize the utility of the tool a modular implementation approach was taken that enables an easy extension of the program (see Appendix B for details on this). The tool implements a workflow that integrates the text mining resources Excerbt and String with the functional analysis tool DAVID. Furthermore, a variety of visualizations and graph measures was implemented in order to support the analysis.

An overview of the implementation of the tool is given in Figure 8.1. As can be seen, the analysis starts with a gene list. This generic format enables the application of the tool to results of various different experiments. On starting the tool, the user is presented with an input window (see Figure 8.2), in which he can load the list from a file (in any format that is supported by DAVID) or simply paste his list of gene identifiers into the input field. In order to properly retrieve the genes from the GO, an organism has to be chosen from a drop-down box. Furthermore, the user can restrict the returned GO terms by requiring a certain level of enrichment and a minimum number of genes that need to fall in a category in order to be considered. The provided gene list is then used to send a query to DAVID's web services. The results of DAVID are in turn put into context by querying the GO Database for the given terms.

Having determined the relevant GO terms the tool then uses either Excerbt or String for providing additional information. The idea behind this is to use the relations between the genes within a GO

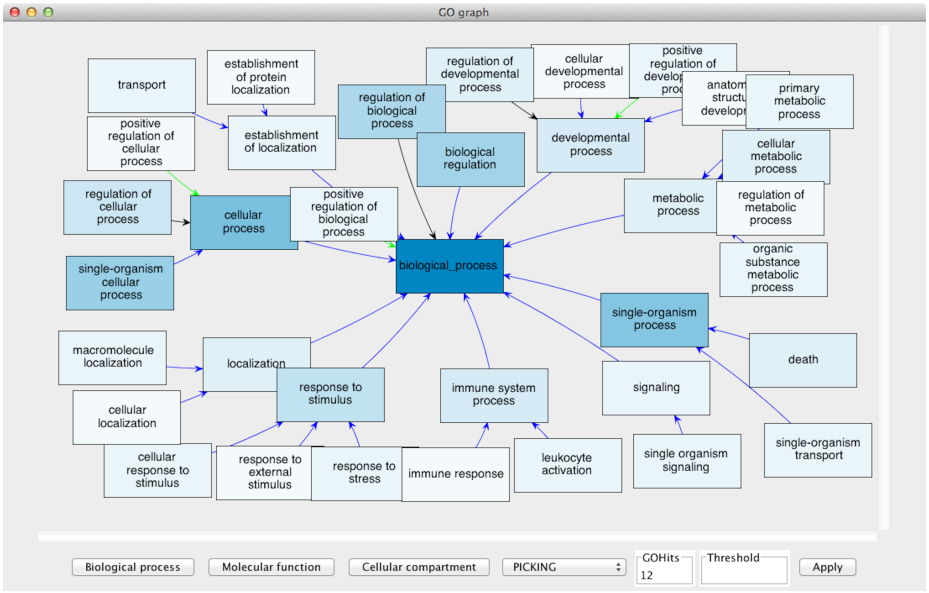


Figure 8.3: Overview window displaying the highest three levels of the GO. Only the enriched GO terms are shown. The color indicates the degree of enrichment with darker terms being more enriched. Picture taken from (Jeske, 2013).

category to offer a more detailed view of the respective functional module. Both Excerbt and String offer text mining results. In contrast to Excerbt, however, String's relations are based on a co-occurrence analysis. Such an analysis typically has a higher recall but a lower precision than more elaborate approaches. Apart from that String also offers relations from other resources besides text mining. It categorizes its additional sources as genomic context, coexpression, and high throughput experiments. Here, genomic context refers to a collection of different prediction methods for functional associations of genes. The other two categories refer to direct results of high-throughput or gene expression experiments correspondingly (von Mering et al., 2003, 2005).

Once the relations are retrieved from the respective resource, they are combined to a graph for each of the enriched GO categories. Furthermore, the following variety of graph measures is calculated for each of the GO terms: the central node, the node with the highest degree, the amount of vertices and edges, the relative size of the largest component of the graph, the diameter, the density and the average clustering coefficient. Finally, all of this information is send to the visualization module of the tool and displayed for the user.

The first thing one is interested in when analyzing the results of an experiment is usually the big picture. In accordance with this, first an overview about the general enrichment of the superclasses is given. An exmaple of this can be seen in Figure 8.3. In this visualization all enriched GO terms of the highest three levels are displayed. The amount of enriched genes within a category is indicated by the color of the node. Darker nodes have more terms, while lighter ones have less. This intuitive representation should help the researcher using the tool to get an immediate impression of which functional categories play an important role in his experiment. In this visualization different relation types are shown in different colors (blue: is_a relation, orange: part_of relation, black: regulates relation, green: positively_regulates

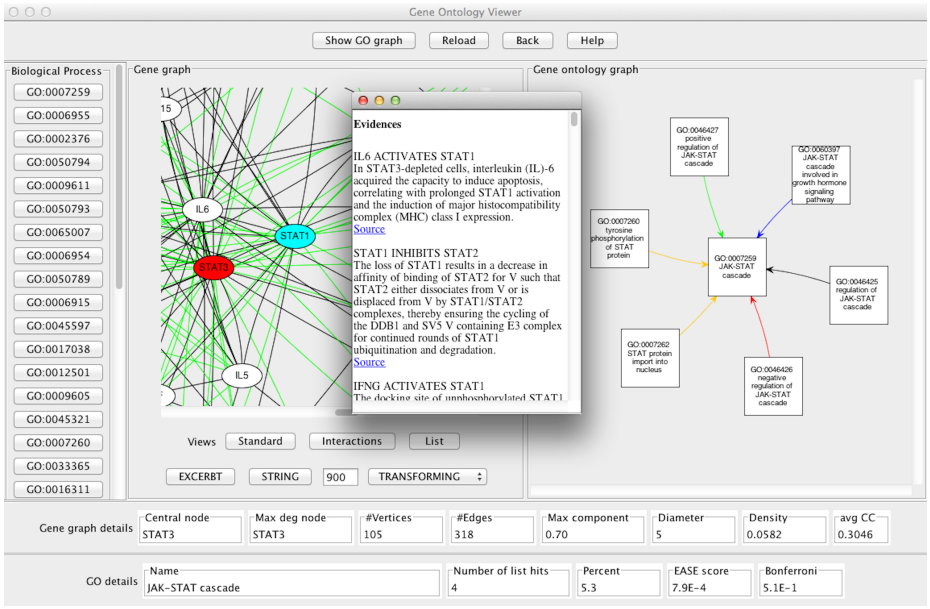


Figure 8.4: Screenshot of the main window of the functional analysis tool. Picture taken from (Jeske, 2013).

relation, red: negatively_regulates relation). Since experimental results can sometimes affect many categories the user is offered the possibility to set a threshold. In this case, only the GO categories containing at least as many genes as asked for are displayed.

After the display of the overview, the user gets the chance to browse through the different GO categories. For this purpose, the user is presented with the graphical user interface (GUI) seen in Figure 8.4. As can be seen, on the left side the enriched GO categories are displayed. The user can choose from them and is presented with the text mining based graphs in the middle of the GUI. Here, in the graph the genes from the input list are displayed in blue. The central genes of the graph as determined by degree and betweeness centrality are displayed in red. It is possible to choose between three different views. In the Standard view all interactions are displayed. Since the resulting graphs can be very complex, however, two additional views are offered that provide a more focussed presentation. In the Interactions view, all unconnected nodes are hidden. In the List view, only the genes from the input list and the central genes are shown.

The connections are again colored like in the overview. The user can click on each node to see a window containing the evidences for the connections of this node as determined by Excerbt. An evidence consists of the relation, the sentence it was extracted from and a link to the paper from which the sentence was taken. Additionally, on the right side the subcategories of the GO term are shown. By clicking on the subcategory the user has the chance to go futher into detail. Apart from that the visualizations are complemented by the graph measures for the respective GO term.

8.3 GO Analysis

The development of the tool was complemented by an analysis of the GO. The Gene Ontology represents the complex interplay of biological systems in a directed acyclic graph. Each abstraction comes at the cost of a restriction of expressiveness. In case of the GO such restrictions include the fact that it is impossible to model temporal or local constraints for relations. In reality certain regulations might only occur under certain conditions. In GO a regulates relation is only included if the regulation occurs always and anywhere. Furthermore, the representation is fully qualitative. The quantitative effects of interactions are not captured and neither are thresholds of concentrations that might trigger interactions. Finally, complex interactions can not be modelled. For example a regulation might only occur if multiple processes occur simultaneously. Such interrelationships that might be modelled with Boolean logic cannot be captured in the GO representation. Finally, the restriction to a directed graph causes the situation that bottom up interactions can be modelled while top down ones cannot. All of these effects are due to the fact that the GO intends to provide a general and static resource. Thus, not all dynamic aspects of biological interactions can be captured and higher levels of detail are traded for a more comprehensive resource.

By modeling the relationships of genes within GO categories, one can gain insights about the structure of the Gene Ontology. Different graph measures might hint towards whether the categories really represent functional modules like they are supposed to do. For this purpose, the number of genes, the density, the clustering coefficient, the diameter, and the average proportion of the largest component of the graph were analyzed with respect to the depth of the GO category within the GO hierarchy. Since there often exist multiple paths from a virtual root node (above the three subparts) to the respective GO term, as a convention the longest of these was chosen. The diameter was calculated on the largest component. If there did not exist any connection in the GO category, this category was left out for the calculation of the diameter and the largest component.

The analysis was performed on 13,288 categories of the Homo sapiens section of GO. Categories containing none, one or more than 1000 genes were excluded from the analysis, since either no meaningful graphs could be build from them or the amount of genes was too large for the String webservice to handle the request. While the categories with zero or one element made up a substantial portion of the GO (60% - 72% of the respective subparts), the exclusion of very large categories was insignificant (0% - 1%). Each of the subparts of GO was considered individually in order to see whether there exist differences between the single ontologies. Furthermore, both the graphs of Excerbt and String were considered individually. This way differences in the sources of the two approaches were accounted for and additionally both text mining resources could be compared. The results of the analysis are given in Figure 8.5.

The first two diagrams (a,b) show the distribution of genes and GO terms with respect to the depth within the ontology. Since higher GO terms subsume the genes of all underlying categories, it was to be expected that the average number of genes within a category decreased with increasing depth. The amount of GO terms on the hand seems to be normally distributed with means around a depth of six (molecular function ontology and cellular compartment ontology) or eight (biological function ontology) respectively. It can be seen that classes with three or less and those with ten or more for

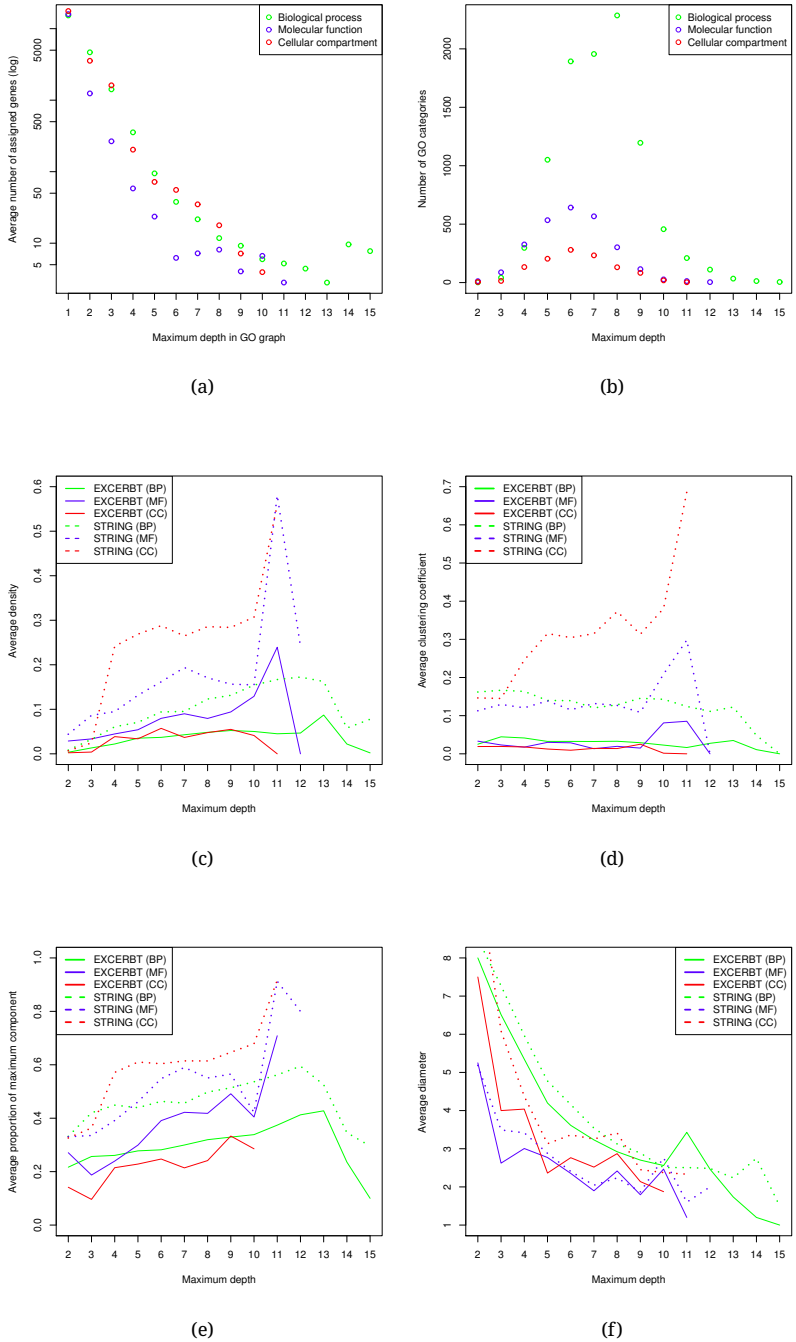


Figure 8.5: Graph measures of GO categories by depth within the GO hierarchy. Pictures taken from (Jeske, 2013).

molecular functions and cellular compartments or 13 and more for biological functions are very rare. Thus, the statistical significance of the graph measures calculated for these is very limited.

The first thing that catches attention when looking at the other plots of the density (c), average clustering coefficient (d), average proportion of the maximum component (e) and the diameter are the peaks around a depth of 10 to 13. As mentioned before, however, the significance of these are small. They can be explained by very few highly connected classes. For example the peak in the Excerbt plot of the density at depth 11 is due to three GO classes that only contain 2-3 genes between all of which Excerbt finds relations. Such classes, however, seem to be outliers that distort the observed results.

Generally it can be seen that the graphs created with relations from String have higher density as well as larger average clustering coefficients and largest components. This observation can be explained by the fact that String finds more connections than Excerbt. Here, String profits from using additional resources besides text mining. This can partly explain the difference which is most striking for the cellular components ontology, probably because cellular components are not that often described in the text of a publication. The difference might, however, also indicate a recall problem of Excerbt. Especially, if one considers that the String results are restricted by a confidence score of 0.900. If this score is decreased to 0.700 the differences become even larger (see Appendix C for details).

The question whether the GO consists of functional modules that become more specific in the lower categories of hierarchy cannot clearly be answered. In specific functional modules one would expect highly connected graphs. The effects that could be observed in the plots are, however, relatively small. The increases in density, clustering coefficient and proportion of the largest component are largest in the regions were there are few GO terms or few genes. Still a similar but weaker tendency can also be seen in the density and the proportion of the largest component. The average clustering coefficient on the other stays fairly constant between a depth of 5 and 9. Additionally, the classes that were not connected at all point towards a questionable ordering in GO categories. Likewise, the proportion of the largest component was rather low. If one (for significance reasons) only considers the GO terms with depth 10 or lower the average largest component only makes up for about 50% - 60% (considering the higher String values). The remaining 40% - 50% are not connected with the main component even on the lower levels. Thus, the genes do not interact with a large amount of genes from the same category. This lack of connection goes hand in hand with a lack of functional relationship of the genes within the category. Based on these findings it seems necessary to bring the ordering of some of the GO category into question. Further research in this direction might prove useful in order to find a system that possibly better represents the functional modularity of the genome.

8.4 Application: mRNA Blood Expression Patterns in Epilepsy Patients Study

In order to test its utility the functional analysis tool was applied to analyze the results of a gene expression study. The chosen study was done by Greiner et al. (Greiner et al., 2013) in 2013. It investigated "mRNA blood expression patterns in new-onset idiopathic pediatric epilepsy" (Greiner et al., 2013). Children with yet untreated epilepsy (37 subjects) were compared to a healthy control

group (28 subjects[8]). Furthermore, the study distinguished between two seizure type subgroups. Here, partial (22 subjects) and generalized seizure epilepsy (15 subjects) were considered individually.

In their functional analysis of the gene list of the partial seizure vs. control group (PvC) , Greiner et al. identified apoptosis, inflammatory defense, and cell motion as important pathways. For the gene list of generalized seizure vs. control group (GvC) , Greiner et al. reported the respiratory chain, mitochondria, and lymphocyte activation pathways as relevant. Greiner et al. obtained their results using DAVID, just like in the text mining supported functional analysis tool. Nevertheless, the functional analysis that was performed in the course of this work could only partially reproduce these results.

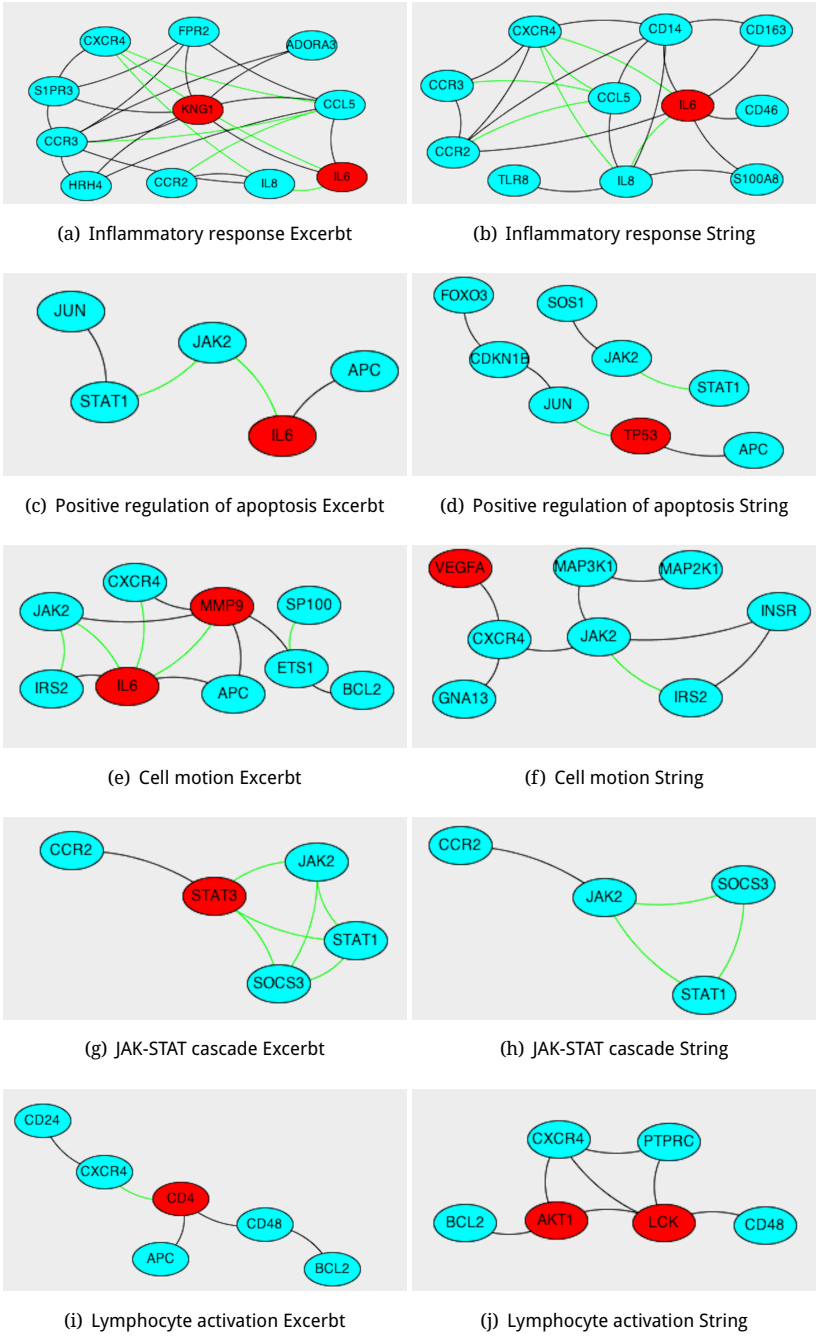
For PvC the reproduced results were similar to that reported by Greiner et al. Several inflammotry and apoptotic processes were significantly enriched. Additionally, cell motion was enriched, but only at a fold change of 1.3 and not as Greiner et al. reported also at a fold change of 1.5. Furthermore, the most enriched GO term JAK-STAT cascade (p-value of 7.9E-4) was not mentioned at all by Greiner et al.

To gain further insights into these GO categories the text mining based graph visualizations of them were created. They can be seen in Figure 8.6. For better readibility the List view of the graphs was chosen. The graphs of the inflammatory response were highly connected. They centered around IL6, which was central both in the Excerbt and the String graph. The importance of IL6 was further supported by the fact that it also appeared prominently in the Excerbt graphs of apoptosis and cell motion. The text mining based analysis supported the involvement of the inflammatory response in partial seizure epilepsy patients. While Greiner et al. could only point to different chemokines and proinflammatory factors, the graphs provided by the integration of text mining showed that and how these are interconnected.

While the interactions of the inflammatory response resembled a highly connected cluster, the connections in apoptosis looked more like a pathway. Both Excerbt and String showed the involvement of the expression regulating genes STAT1 and JUN. Furthermore, JAK2 appeared in both graphs. Considering that the JAK-STAT cascade was the most enriched term, this could indicate that the enrichment of apoptosis might be a side effect of the involvement of this cascade. In this connection, it should also be pointed out that further genes found in the JAK-STAT graphs could also be seen in those of the inflammatory response (CCR2) and cell motion (JAK2). Thus, the functional analysis pointed towards the JAK-STAT cascade as the most relevant GO category. Additionally, it showed how other categories like the positive regulation of apoptosis, cell motion and the inflammatory response are connected with it. This goes beyond the comparatively shallow analysis of simply listing enriched GO terms.

For GvC, the results of Greiner et al. could not be reproduced. Instead of the respiratory chain, mitochondria and lymphocyte activation among the most enriched categories were GO terms concerning the cell cycle and nuclear import. Mitochondrial membrane and respiratory chain at least for a fold change of 1.3 got p-values of 1.1E-3 and 1.3E-3 repectively. Lymphocyte activation was not enriched for GvC but instead showed an enrichment for PvC at a fold change of 1.3. The poor reproducibility of these results made it impossible to refine the interpretations of Greiner et al. Still a look into some relevant categories is given in Appendix D.

[8] It should be pointed out that the size of the control group is considerably too small which might be one explanation of the comparatively poor results of the study.



(a) Inflammatory response Excerbt

(b) Inflammatory response String

(c) Positive regulation of apoptosis Excerbt

(d) Positive regulation of apoptosis String

(e) Cell motion Excerbt

(f) Cell motion String

(g) JAK-STAT cascade Excerbt

(h) JAK-STAT cascade String

(i) Lymphocyte activation Excerbt

(j) Lymphocyte activation String

Figure 8.6: Text mining based graphs of GO terms relevant for partial seizure epilepsy patients. Pictures taken from (Jeske, 2013).

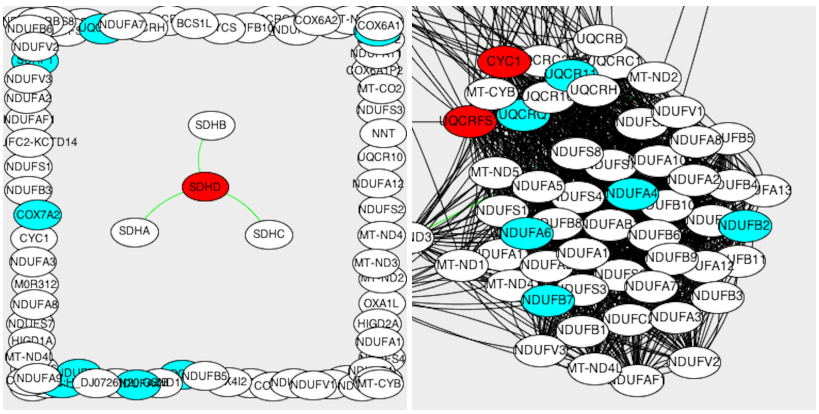
Beside the insights one could gain about GO categories, the proposed functional analysis tool also revealed some interesting aspects of the used text mining resources. The first aspect that drew attention in this respect was the unreliability of the availability of the webservices of the two tools. This was more problematic for Excerbt but was partially resolved by Benedikt Wachinger during the work on the analysis tool. In addition to that Excerbt's unusual way of handling entities became apparent. Instead of focussing on biological entities and adding synonyms to them, Excerbt has entities for every different spelling and synonym of a term. These different spellings are in turn connected via synonym relations. However, these are not automatically resolved, when queried. Thus, e.g. if one sends a query to Excerbt for results of 'brca1', Excerbt does not return results for 'BRCA-1' or 'BRCA1'. One could circumvent this problem by first querying for the synonyms of a term and then subsequently including all results into the query. This, however, causes new problems since Excerbt often failed to execute too complicated queries. Furthermore, there are inconsistencies in the synonym module of Excerbt. The syonym relations are not symmetrical, how it would be expected. For example '20S Proteasme' is returned as a syonym of 'Macropain' but 'Macropain' is not returned as a synonym of '20S Proteasome'. Similarly, for 'phi-1' 4 synonyms are returned, for its synonym 'ppp1r14b' 13 synonyms are returned.

Finally, the graph visualizations of the two text mining resources revealed differences in the respective coverages of relations between genes. This might have become most obvious in the visualization of the respiratory chain seen in Figure 8.7. As can be seen, String returns a highly connected cluster of 1417 relations, while Excerbt only finds three interactions. The most likely explanation for the differences in the two graphs is the fact that respiratory chain comprises many proteins that aggregate in complexes like the NADH dehydrogenase. Such complexes are rarely explicitely described in literature. Thus, text mining approaches have no chance of extracting these relations. Here, String profits from the fact that it includes several resources. The respiratory chain is also a good example where the functional analysis including resources like String can support the interpretation of gene expression results. Greiner et al. claimed that the role of mitochondrial gene expression in epilepsy was unknown. Closer examination of the String graph of the respiratory chain, however, pointed towards an involvement of the NADH dehydrogenase and the cytochrom c reductase, since various genes of these were differentially expressed.

8.5 Conclusion

Especially, the last figure (8.7) obviously demonstrated the need of knowledge integration. While the purely text mining based Excerbt could only find three connections in the GO class "respiratory chain", the visualization of String that is additionally based on other resources is a hairy ball. For a comprehensive picture of the current knowledge about a topic all relevant resources need to be considered.

While the integration of the results of text mining with prior knowledge is rather intuitive, there might also be value in integrating a priori knowledge in the text mining process itself. Since communication is optimized for efficient comunication, already known facts are only briefly referenced or completely left out. Thus, in order to get a comprehensive understanding of a text a text mining systems requires similar background information like the author of the text. For example, knowing that certain patterns are well-established might help to better interpret ambigue, unclear or uncertain formulations.



(a) Excerbt (b) String

Figure 8.7: Text mining based graphs of GO term respiratory chain for generalized seizure epilepsy patients. Pictures taken from (Jeske, 2013).

8.6 Related Work

Table 8.1 shows related approaches to the integration of text mining results into functional analyses.

Table 8.1: Text mining-based functional analysis: Related work

Authors	Approach
Medina et al. (2010)	Gene expression analysis suite containing text mining and GO categorization
Gotz et al. (2008)	Functional annotation of DNA or protein sequences using GO-graph visualization
Hur et al. (2009)	Text mining of continuous texts with subsequent functional analysis
Al-Shahrour et al. (2007)	Text mining for additional functional analysis
Tiffin et al. (2005)	Integration of text mining and data mining of gene expression data
Scherf et al. (2005)	Integration of genomic analysis into text mining to improve performance
Lussier et al. (2006)	Text mining for providing phenotype context information for GO annotations

9

Towards a Supersemantic Analysis I - Shallow SRL

9.1 Excerbt Restrictions

The supersemantic applications presented in the last chapters were either implemented from scratch or based on the Excerbt text mining system that was developed at the IBIS (the only exception to this is the anaphora resolution system presented in chapter 6). During the work on these applications many shortcomings of Excerbt became obvious. Some of these impaired the results of the developed algorithms and tools. All of them, however, have an impact on the quality of the overall results of a possible supersemantic text mining system that is based on Excerbt. The apparent problems of Excerbt are the following:

A very coarse grained ontology

Ontologies are commonly hierarchically organized systems with many intermediate levels. The entries in the different levels are connected with is_a, part_of or even more specific relations. For example, in the Gene Ontology, the proteasome complex is defined as a protein complex which in turn is a cellular component. Likewise, the nuclear proteasome complex is a proteasome complex. Such a hierarchical representation allows to simplify searches. For example one can easily look for all kinds of protein complexes without having to list each single one.

Furthermore, it allows to choose the level of detail that seems most appropriate for the given task. Exemplarily, for the creation of the Negatome protein complexes and mutations were excluded. Excerbt, however, is based on an ontology that is too coarse grained to distinguish them from other proteins. Instead, Excerbt only has one big gene/protein category in which all of these fall. In the Negatome example, this resulted in the difference between the sample evaluation and the final acceptance rate.

While during the sample evaluation all kinds of interactions involving proteins were accepted, in the final annoation of the Negatome mutations and complexes were excluded. This lead to a drop in accurracy from Excerbt of about 15%-20%. In the highest confidence interval (according to the confidence score) this value even reached over 40%.

In addition to that, Excerbt's ontology does not contain any subcategories. Unfortunately, this significantly increases the amount of irrelevant results returned for queries in many application fields and even completely disqualifies Excerbt in others.

A lack of mappings to existing resources

A second shortcoming of the ontology used by Excerbt is the fact that it does not map to other resources. This is especially unnecessary since it was originally largely based on other ontologies. This way it is hard to keep track of changes in the underlying ontologies which in turn explains why the ontology is not frequently updated as new versions of the source ontologies are created. A comprehensive system would try to automate this process as much as possible and would provide mappings. This shortcoming has also become a nuisance in the creation of the Negatome. Since the format in which the Negatome is published requires gene identifiers, these had to be manually added by the curators. This additional work could have been diminished if Excerbt would provide the corresponding identifiers.

An incoherent entity system

Quite a few things about the entities in Excerbt are counterintuitive. Since Excerbt's ontology is based on multiple resources, it contains the same things multiple times. Instead of mapping these entities onto each other, however, each entry is kept and treated individually. Furthermore, even the synonyms taken from the same resources are treated individually. Thus, Excerbt returns thousands of results for p53 but none for prac which is listed as a synonym of p53 in EntrezGene (one of the resources Excerbt's ontology is based on).

Instead the user has to actively select all listed synonyms in order to get the results. The synonym system, however, is inconsistent itself. Synonyms are by definition bidirectional. Yet, Excerbt does not treat them that way. Thus, for p53 there are 87 synonyms listed. For prac, on the other, Excerbt only offers 12 synonyms.

Additionally, there seem to be synonyms missing - even some very common ones. The term Alzheimer's e.g. is only considered as a pathway but not as a disease/phenotype.

The necessity of syntactical post-processing

Syntactic processing is very time-consuming. For this reason, the design choice to avoid it was taken when developing Excerbt. Instead Excerbt was based on Senna - a tool that largely circumvents syntactic processing and instead immediately provides a semantic interpretation of an input sentence. Senna is trained on the Propbank corpus and correspondingly classifies semantic roles as described in Propbank. Unfortunately, due to the semantic role definitions of Propbank these roles are often not

suited for the immediate extraction of biological events. Instead syntactic post-processing would be required to avoid an accumulation of false positive results.

The problem of the role definitions was e.g. pointed out by Robert Strache (Strache, 2012) in his Master thesis by providing the following example:

> "At the light microscope level, brush cells can be identified by antibodies against the actin filament crosslinking proteins villin and fimbrin that not only stain the apical tuft of microvilli and their rootlets, but also label projections emanating from the basolateral surface of these cells."

For this sentence, Senna returns, among others, the following predicate-argument-structure:

PRED:	"identified"
ARG0:	"by antibodies against the actin filament crosslinking proteins villin and fimbrin that not only stain the apical tuft of microvilli and their rootlets"
ARG1:	"brush cells"
ARGM-MOD:	"can"
ARGM-LOC:	"At the light microscope level"

The very simple event extraction step of the original Excerbt implementation simply looked in the ARG0 and ARG1 for biological entities. This lead to the extraction of erroneous events, like "rootlets identified brush cells" in the given example. This approach was improved by a heuristic implemented by Robert Strache in his Master thesis. According to this heuristic, only entities that occurred before the first verb within the argument were used for event extraction. This restriction is meant to deal with cases where subclauses occur in arguments. Thus, in the given example, 'rootlets' is discarded since it occurs after the first verb ("crosslinking").

The heuristic tries to make up for the lack of a proper syntactic analysis, but is only able to cover certain special cases. While it succeeds at discarding objects of subclauses within an argument, it fails to discard the subjects of the subclauses. For example in the sentence "Alzheimer's is caused by brain cell death that Smithee et al. attribute to plaques." Senna would tag the end of the sentence beginning with "by" as ARG0, the heuristic would eliminate everything after "attribute" but "Smithee et al." would still be a valid argument. Furthermore, genitive expressions and word combinations are not handled properly. For example, the sentence "Liver diseases cause depressions" would produce an event cause(Liver,depressions) or the sentence "Bacteria can cause heart infections" would produce an event cause(Bacteria, heart). A proper linguistic analysis would determine the head word of such expressions in order to get the right results.

Senna is a blackbox

Senna is an external machine-learning tool. The problem that arises when using machine learning models is that for each alternation in ones problem definition or solution strategy one has to train a new model. Furthermore, one requires the corresponding training samples that properly represent the problem one tries to solve. The acquisition of such training corpora is a very time-consuming process. Thus, if one e.g. wanted to correct the annotation scheme of Senna in order to avoid annotating whole

subclauses into arguments, one would first need to annotate a new corpus or use a completely new tool that accomplishes the task and can be used as post-processing.

Additionally, machine learning models like the deep neural net used by Senna are a black box whose information is not in any kind of human readable format. Thus, one is very restricted in identifying and fixing problems in the classification. Finally, Senna is an external tool whose source code is not publicly available. Because of this it is not possible to modify the classification algorithm. Even if one only wants to make small changes post-processing or reimplementation are the only choices. This situation strongly restricts the possibility of including improvements into the Excerbt workflow.

A need for verbs to extract events

Excerbt's event extraction is exclusively based on Senna. Senna, in turn, extracts predicate-argument-structures that are always rooted in verbs. Verbs, however, are not the only way events can be described. This approach is very restricting and is a main source of Excerbt's recall problems. In order to give an idea of the extent of this restriction, Table 9.1 gives an overview over forms of expressions missed by Excerbt.

Table 9.1: Utterances that describe an event without using a verb.

Type	Example
Nominalization of predicate	The **activation** of Bax by p53
Argument nominalization	P53 is **a regulator** of Bax.
Adjectives	The **p53-induced** activation of Bax
Ellipsis	P53 activated Bax, **but not** lrrk2
Multiple clauses	Bax is activated, **when** p53 occurs in the cell.

The sample evaluation performed in the course of the Negatome analysis revealed that about 25% of relations are not found because they are formulated in one of these ways. Thus, the restriction to verb events was the biggest recall problem in this, however, limited evaluation. It was followed by a lack of anaphora resolution (20%) and a lack of nested events (5%). The last of these again can be attributed to the restricted approach taken by Excerbt.

A restriction to simple events

In the Negatome sample evaluation the focus was on non-interacting proteins. These are usually described in simple events. In order to get a more comprehensive picture of the biological knowledge described in a publication, one needs a more powerful framework that is able to extract and represent nested events. Already in the small Negatome evaluation one non-interaction was missed due to the lacking possibility to extract nested structures. For other tasks this is even more important. Nested events can be used to express causal chains of events but often also direct relations are described in a nested way. Take for example the adjectives example from Table 9.1. Even though the direct interaction of p53 and Bax is described, linguistically it is a nested event that could be written as induction(p53, activation(Bax)). Such formulations occur often in publications and are thus an essential

part of contemporary text mining evaluations like the BioNLP shared task 2013. Excerbt ignores such events, which further limits its applicability.

A lacking sanity-check for events

Well defined events impose certain restrictions on what kind of entities are allowed as their players. For example, a gene expression event needs by definition a gene as theme (ARG1) or a protein binding event needs at least one argument that is a protein and the second argument has to be some kind of substance or another protein. A comprehensive event extraction system should take these restrictions into account in order to improve the precision of its results.

Unfortunately, Excerbt is missing such a sanity check. Events in Excerbt are only defined by the verbs by which they are identified. All biological entities that are found in predicate-argument-structures are combined to such events independent of whether this would make sense. The consequence of this are results like the following:

- **Sentence**: “For example, report of a fall would trigger a home safety assessment, whereas loss of a loved one or pet would trigger a depression evaluation by the social worker.”

  Entities: pet $\in$ Gene, social worker $\in$ Person

  Event: Activation(pet, social worker)

- **Sentence**: “’We can extrapolate from the United States to a degree,’ says Ferguson, ’but there are too many variables to judge accurately.’ The United States has a lot of automobiles, and compared to many other countries, Americans tend to build more (and wider) roads, more (and bigger) parking lots, more (and more expensive) shopping centers, and larger houses (with accompanying larger roofs).”

  Entities: parking lot $\in$ Environmental factor, countries $\in$ Geolocation

  Event: Binding(parking lot, countries)

- **Sentence**: “OBJECTIVES: After more than 10 years’ experience in France, the French Foot Surgery Association (Association francaise de chirurgie du pied [AFCP]) presents an update on mobile-bearing ankle prostheses, based on a multicenter study.”

  Entities: France $\in$ Geolocation, bears $\in$ Species

  Event: Expression(France, bears)

- **Sentence**: “The library was initially depleted of phages recognized by naive mouse serum by 3 sequential pannings of the library with immobilized serum of non-immunized mice.”

  Entities: library $\in$ Environmental factor, mouse $\in$ Species

  Event: Inhibition(mouse, library)

These events should have been defined on the entities that they can actually occur with. Instead, geolocations can trigger gene expression events, libraries can be inhibited and countries bind to

parking lots. Besides the lacking sanity check, these example of course also reveal other shortcomings of Excerbt. For example, in the last sentence the fact that 'mouse' is found as an argument is due to the insufficient syntactic processing. A proper syntactic analysis would have detected serum as the head word of the chunk. Furthermore, in the first sentence, the lacking detection and processing of nominalizations and nested events is responsible for the wrong event.

A lack of a comprehensive evaluation

One possible reason this widespread range of problems was not addressed yet, might be the lack of a comprehensive evaluation that could have brought them to light. The only evaluation that was conducted, however, could only give limited insights. Robert Strache (Strache, 2012) evaluate Excerbt on data from BioInfer (Pyysalo et al., 2007b), IntAct (Kerrien et al., 2012) and BioGRID (Stark et al., 2006) in his Master Thesis.

The BioInfer evaluation resulted in a low F-measure of 0.113. However, even this very low value is sugarcoated since certain events were taken out of the evaluation beforehand. Whenever biological entities were described in a more complicated way that overstrained Excerbt's simplistic named entity recognition, they were left out. Thus, the evaluation only focussed on the event extraction but not the named entity recognition necessary to that end. The meaningfulness of the BioInfer evaluation is furthermore doubtful because of the shortcomings of the BioInfer corpus reported by Robert Strache.

The evaluations on the IntAct and BioGRID data were no evaluations on annotated corpora but on databases containing known protein-protein interactions. Text mining systems are commonly evaluated by measuring the F-measure of the instances of found biological events within texts. Thus, the evaluation on these two resources cannot act as a valuable comparison with other text mining tools. Furthermore, the obtained results of F-measures of $4.76 * 10^{-3}$ and $9.05 * 10^{-3}$ respectively were alarmingly low. Here, the F-measures were calculated on relations level, which is unorthodox and complicates the comparability. Furthermore, both resources might lack protein-protein interactions (PPIs) that are described in the literature and the other way around PPIs within those resources might not be mentioned in any of the texts analyzed by Excerbt. Thus, this second evaluation is even less comprehensive than the first one.

Summing up one can conclude that Excerbt is lacking a proper evaluation. The restricted obtained results, however, seem to indicate a huge quality problem. While the recall problem was already mentioned in this chapter, the precision problem might yet be worse. In all three of the evaluations the precision was lower than the recall. Thus, a comprehensive evaluation might only be the first step in an attempt to create a trustworthy text mining system.

9.2 Shallow Semantic Role Labeling

This list of shortcomings suggests that the development of a more comprehensive text mining system is necessary. Such an endeavour, however, requires a huge effort. By way of illustration, BioContext, currently probably the best comparison for a contextualized Excerbt, was developed over three years (though some components are even older) at a national center for text mining. The different

contextualization, text analysis and event extraction modules were published separately by six different first authors in nine different publications (this only includes the publications of the researchers and their collaboration partners, modules that were from other research groups are not counted) (Gerner et al., 2012). This can illustrate to some extend the necessary workload of developing a modern text mining system. Implementing a complete one along the lines within the course of my thesis is therefore unrealistic. Nevertheless, the insights gained by the work on the different supersemantic methods can be put to use in two ways. First, a concept and roadmap on how a modern text mining system that avoids the mistakes of its predecessors should look like can be provided. And secondly, prototypes that illustrate at least certain relevant aspects of such a system can be implemented. Both of these were realized in the course of this work. The concept of a modern supersemantic text mining system was already presented in section 2.5. Two prototypes will be presented in the rest of this and the following chapter.

The first prototype is called Shallow SRL. The idea behind it is to tackle the lacking syntactical processing at least partially and to implement a system that is less of a black box and instead can be more easily extended and changed. The main problem when including a syntactic analysis is a loss of efficiency. BioContext that included multiple sentence analysis tools took six months to process Medline (Gerner et al., 2012) and Thorsten Barnickel examined different tools for syntactical processing in the course of his doctoral thesis with the result that none was suitable for an efficient application to the large amount of biomedical publications that currently exist (Barnickel, 2009). In order to include the necessary syntactic information without loosing efficiancy, in Shallow SRL I steered a middle course.

9.3 Approach

Instead of applying a full parsing approach only shallow parsing, or chunking how it is often referred to, was used. In syntactic analyses, the words of a sentence are ordered in a hierarchy of groups of related utterances. For example, articles and the corresponding nouns form noun phrases, verbs and noun phrases form verb phrases and noun and verb phrases can be combined to sentences. In a shallow parsing approach, this procedure is simplified by leaving out the hierarchical structure and instead only chunking the lowest level of the syntactic analysis. This level includes the chunks described in Table 9.2.

Shallow parsing is very efficient compared to full parsing. Thus, it is a way of obtaining some syntactic information without suffering from the processing time drawbacks of other parsing approaches. In Shallow SRL, this approach is complemented with a second heuristic in order to get additional syntactic information in an efficient way. The second heuristic is meant to detect clause boundaries within a sentence. This was motivated by the problems observed in Excerbt. By detecting the clauses of a sentence before extracting events, one can avoid having whole clauses within arguments. The chunks within a clause are then labeled by a machine learning algorithm that was (like Senna) trained on the Propbank corpus (Palmer et al., 2005).

An overview over the sentence processing pipeline is given in Figure 9.1. The sentences themselves are detected by the sentence detector of the OpenNLP Java library (The Apache Software Foundation, 2010). In the first step of the pipeline, the following preprocessings are performed on the sentence:

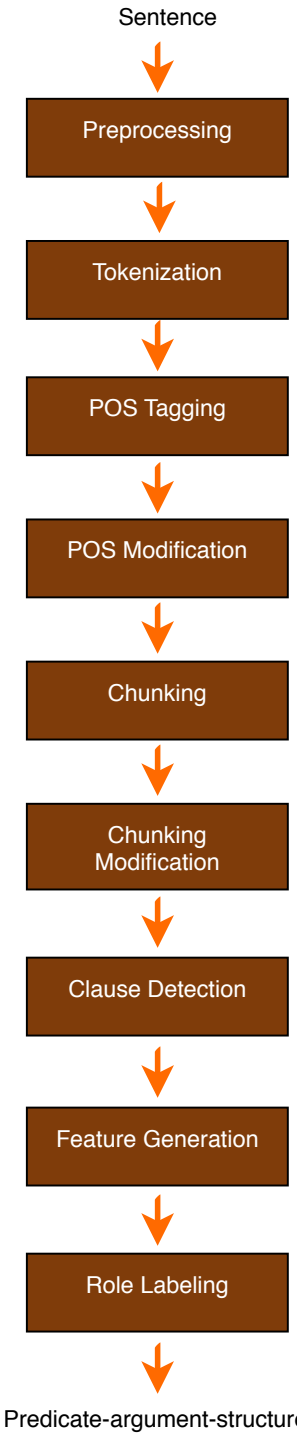


Figure 9.1: Shallow SRL processing pipeline.

Table 9.2: Chunk types of the OpenNLP chunker. The type set is based on CoNLL-2000 shared task (Sang and Buchholz, 2000). The examples are mostly taken from Sang and Buchholz (Sang and Buchholz, 2000).

Type	Label	Example/description
Noun phrase	NP	“the most volatile form”
Verb phrase	VP	“has got”
Adjective phrase	ADJP	“old”
Adverb phrase	ADVP	“earlier”
Prepositional phrase	PP	“such as”
Subordinate clause	SBAR	“even though”
Particles	PRT	“on and off”
Conjunction phrase	CONJP	“as well as”
Interjection	INTJ	“good grief!”
Punctuation mark	O	“,”
List marker	LST	All kinds of listings
Unlike coordinated phrase	UCP	Collective term for everything else

- Line breaks and tabs are removed.
- Everything between any kind of brackets is cut out and treated as a sentence on its own.
- Everything between two hyphens is cut out and treated as a sentence on its own.
- The sentence is split at the following characters: ’;’, ’:’ and ’-’. Both parts are treated as separate sentences.
- Quotation marks are changed to match the format of Propbank.

The main intent of these preprocessings is to filter out additional sentences that are embedded in the one under investigation. Furthermore, the complexity of the sentence structure should be reduced. The preprocessing is followed by two typical natural language processing steps. The sentence is first tokenized and the tokens are subsequently POS tagged. Both steps are again performed using the implementations of the OpenNLP package. Since the OpenNLP tools were trained on financial text data they sometimes experience problems when confronted with entities from the biological domain. Especially, gene aliases are occasionally assigned a wrong POS tag. To counteract this behavior to some extent a rule-based post-processing of the POS tags is performed. Here, words that contain at least one digit but are not tagged as number (CD) or nouns are relabeled as nouns.

The modified POS tags are then used as input for the chunker. Again, the OpenNLP implementation is used for this step and again the results of the tool are modified according to a set of rules. The applied chunking modification rules are the following:

- Chunks containing two possible NPs seperated by a ’and’ are split into three chunks (the two NPs and the ’and’).
- Chunks are split at commas, brackets and hyphens. The separation characters and the rest of the chunk before and after it are again turned into separate chunks.

- Two NPs and the chunk between them are combined if the intermediate chunk contained only the expressions ' of ' or ' 's '.
- Two NPs and the chunk between them are combined if the intermediate chunk contained only of a single apostrophe and the first NP ends with a 's'.
- An enumeration detection heuristic is applied that tries to combine chunks of the same type that are connected with commas and the coordinating conjunctions 'and' and 'or'.

The objective of the chunking post-processing steps is to get the chunks in a form that most precisely resembles the arguments that should be extracted in the role labeling step. Thus, genitive constructions and enumerations are combined into single chunks. For the enumerations, a heuristic was constructed. The problem when trying to detect enumerations in a context-free manner is that they can look the same as the point of contact of two clauses that are combined by an 'and' or a 'or'. Take for example the utterance 'p53 and lrrk2'. This looks like an enumeration on first sight. It could, however occur in a sentence like the following:

> Bax activates **p53 and lrrk2** inhibits it.

In this case, the clauses are enumerated and not the noun phrases. There exist several additional formulations where enumerations look similar to other utterances. Therefore, the developed heuristic goes beyond merely combining utterances that look like enumerations at first. For this, the range of the enumeration and the amount of verbs within the respective clauses are considered. Only utterances that span over more than two entities are considerd enumerations. If this is not the case, the heuristic considers the two clauses that would be created if the enumeration is not created. If both of the clauses contain a verb, no enumeration is created, otherwise it is. While this procedure is not perfect, it at least provides an improvement over ignoring enumerations.

Afterwards, the clause boundaries are detected by a set of rules: Commas and a set of unambigue function words that commonly indicate subordinate clauses are considered as clause boundaries. The following function words were used: although, because, but, how, if, that, though, what, when, where, whether, which, while, who, whom, why. Additional ones, like 'since' might be included in the future. However, here a disambiguation is required in order to distinguish occurrences of 'since' that begin a subordinate clause from those that only begin a prepositional phrase like in 'since 1990'. Furthermore, infinitives were considered as indicating clause boundaries if they were not in a verb chunk with a finite verb. This was done in order to create subordinate clauses with at most one verb, thus, reducing the argument search space. In infinitive constructions, one word is often used as argument for two predicates. Take for example the following sentence:

> Bill told Bob to go to Beth.

Here, Bob is the ARG1 of the predicate 'tell' and the ARG0 of 'go' at the same time. Such constructions are a way of shortening expressions. When extracting events from them this has to be considered. Thus, the noun phrase preceding the infinitive clause needs to be considered as the missing argument in the infinitive clause. This feature was not yet implemented, since it was irrelevant for the tool Shallow SRL was applied in. In the future, however, this should be included, if Shallow SRL is used for generalized event extraction. In the English language, infinitive clauses and some subordinate clauses are not surrounded with commas. Thus, in these cases the end of the clause was predicted by a heuristic. The

clause was considered to end when another verb occurred. However, if the other verb was preceded by a noun this noun was not considered part of the subordinate clause but instead subject of the verb. This was done to avoid the problems described in the 'The necessity of syntactical post-processing' section.

Once the clauses were detected, they are labeled with a specific type. The following types were considered: main clause, relative clause, causal clause, contrary clause, conditional clause, other subclause, ellipsis, said-clause, that-clause, infinitive clause, gerund clause, temporal clause, location clause, temporal insertion, manner clause, other verb modification. This is a mixture of syntactical and semantical types. Where possible, e.g. in the case of tempus or location, the semantic interpretation of the clauses was already assigned. If this was not obvious, a syntactic type, like relative clause, was assigned. Main clauses that were interrupted by subordinate clauses (e.g. by the insertion of a relative clause after the subject) were merged to be treated like one clause.

Having detected and classified the clauses, now features for the machine learning algorithm were created. Each chunk was turned into a feature vector for each verb within its clause (this should be maximally one if the clause heuristic works perfectly). Table 9.3 gives an overview over the used features.

Table 9.3: Overview over the features used by the support vector machine in Shallow SRL.

Feature
Distance of chunk in question (CIQ) to verb
Word embeddings of head words in five chunk window around CIQ
Word embeddings of head words in five chunk window around verb
Voice of the verb
Amount of verb chunks between CIQ and verb
Amount of prepositional phrases between CIQ and verb
Amount of other chunks (0) between CIQ and verb
POS types of head words in five chunk window around CIQ
POS types of head words in five chunk window around verb
Amount of words in CIQ
Chunk types in a five chunk window around CIQ
Chunk types in a five chunk window around verb
Amount of verbs in the sentence
Whether CIQ was preceded by 'who' or 'which'
Whether CIQ was preceded by 'whom' or 'whose'
Whether CIQ was preceded by 'that'
Whether the first word in CIQ is from a list of temporal terms
Whether the last word in CIQ is from another list of temporal terms
Whether the sentence contains subordinate clauses
Whether the chunk is a placeholder for a subordinate chunk
If it is a placeholder, the type of the subordinate clause
Whether the current clause is a placeholder in another sentence

The used word embeddings were obtained from Collobert and Weston as published on the metaoptimze website (Turian et al., 2010a). Word embeddings are vector representations of words. Like the term

vectors described in chapter 7, they describe a word space model. Collobert and Weston created theirs by training a deep neural net. The distance to verb feature was bound to values of up to +/- 5 to avoid data sparseness. Likewise, the amount of intermediate verb chunks was bound to three, the amount of words within the chunk in question to four, and the amount of verbs in the sentence to seven. Two lists of temporal terms were created to check for temporal arguments. The first list contained words that typically stand at the beginning of a temporal expression, like 'after' in 'after the game'. Likewise, the second list contained words that typically stand at the end of a temporal expression, like 'ago' in 'three years ago'. In nested events, whole subordinate clauses can be arguments. To account for such situations, subordinate clauses can - depending on their type - be treated as chunks in their respective superordinate clauses. This allows them to be labeled just the same ways as normal chunks. The last three features in Table 9.3 refer to such placeholder chunks.

The features were used to train, test and apply a linear support vector machine. The Java-implementation of LibLinear (Waldvogel, 2013) was used for this. The classification was performed on chunk level with the restriction that only chunks within the same clause as the respective verb were classified. Since verb chunks can contain ArgM-MOD and ArgM-NEG arguments, these were extracted rule-based in a post-processing step. For this, modal verbs and the negation terms 'not', 'n't' and 'never' were extracted and labeled accordingly. The final predicate-argument structure was then used analogously to Senna in Excerbt to get an improved event extraction.

9.4 Evaluation & Application

Shallow SRL was evaluated on section 23 of the Propbank corpus. This section is traditionally used for evaluation of semantic role labeling systems. Thus, the comparability with other systems is given. The evaluation was performed on chunk level in order to match the problem representation of Shallow SRL. Furthermore, the systems performance was evaluated within the clause boundaries. Here, a recall of 0.74, a precision of 0.7 and a f-measure of 0.72 was reached. This value is comparable but slightly lower than the reported f-measure of 0.75 of Senna on section 23 of the same corpus. Additionally, the efficiency of the two approaches was compared. Shallow SRL needed on average 7.68 miliseconds to process one sentence on a Mac Book Air with a 1.7 GHz Intel Core i5 processor with 4 GB RAM. Senna, on the other hand, took 98.71 miliseconds per sentence on the same hardware. Thus, Shallow SRL has a significant efficiency advantage.

Furthermore, Shallow SRL was applied to a practical problem of text mining. Dictionaries and glossaries are useful tools in every specialized domain. They can be used as a reference system as well as to look up unknown terms. An overview over related work on this is given in Table 9.4. As one can see, multiple systems have already been proposed, but within the biomedical domain for English texts to my knowledge only the Definder system of Klavans and Murescan (Klavans and Muresan, 2001) exists. The construction of such resources is very time- and labor-consuming. Furthermore, languages are dynamical. New terms and concepts emerge on a regular basis which quickly lets the existing resources become outdated. As a consequence of this, very specialized domains often lack appropriate up-to-date dictionaries. In order to overcome this shortcoming, automated approaches to extract definitions from unstructured text corpora need to be developed. In the course of this thesis, Shallow SRL was adapted for the specialized problem of definition extraction in a tool called DefineTHAT.

In order to automatically extract definitions, two central problems need to be solved. First, sentences containing definitions need to be identified and secondly the term that is defined needs to be extracted. To accomplish the first task, Shallow SRL was combined with a rule-based filter that determines whether the sentence under investigation contains a definition. The identification of the defined term then boiled down to identifying the subject of the corresponding statement. The filter identified a sentence as containing a definition if one of the following conditions was fulfilled:

- The verb is a form of “to be”, there exists a noun chunk in front of the verb, the word before the verb is not “there” and the verb is followed by an article.
- The verb is a plural form of “to be”, there exists a noun chunk in front of the verb, the word before the verb is not “there” and the verb is followed by an adjective and a noun.
- The verb is a form of “to be”, there exists a noun chunk in front of the verb, the word before the verb is not “there” and the verb is followed by a noun.
- The verb is a form of “to be”, there exists a noun chunk in front of the verb, the word before the verb is not “there” and the verb is followed by an adverb and an article.

If a sentence was detected as containing a definition, the term that is defined was identified by role labeling the sentence. Then the ARG0 and ARG1 roles with respect to the “to be”-verb were considered. Since Propbank is not annotated for “to be” there is no reliable destinction between the two. Thus, they are treated as equally likely candidates. A sample evaluation on the considered corpus revealed, however, that the defined term is usually (in around 95% of the cases) in the subject of the sentence. Consequently, the ARG0 or ARG1 that is in this position is taken.

The approach was evaluated on a part of the Wesbury Lab Wikipedia corpus (Shaoul and Westbury, 2010). The sentences were first automatically filtered to only include sentences that were detected by the rule scheme described above. Afterwards they were manually inspected to correct sentence parsing errors in the corpus and to label the subject of all “to be” verbs within it. In total, 451 sentences were labelled manually. Testing the system on a data set from domains independent of the data the classifier was trained on should increase the expressiveness of the obtained result with respect to its generalisability. The statistical values were calculate word-based and on the complete sentence to facilitate comparability. The classifier reached a precision of 0.93, a recall of 0.94 and a f-measure of 0.94 on word level. All in all, 89.14% of the sentences were classified completely correctly. The classifier was applied to a corpus, consisting of more than 120 million sentences. This yielded approximately 11.1 million definitions. Providing the results as a website and an Android-App is currently in preparation.

9.5 Related Work

Table 9.4 gives an overview over other approaches to definition extraction. One can see that the focus of most of the systems is on e-learning and question answering. The picture is more heterogeneous for the languages of the systems. For a wide variety of languages, definition detection systems have been developed.

Table 9.4: Related work on automatic definition detection.

Authors	Approach	Language	Domain
(Klavans and Muresan, 2001)	Rule-based	English	Medicine
(Gaudio and Branco, 2007)	Rule-based	Portugese	E-Learning
(Navigli and Velardi, 2010)	Word class lattices	English	Various domains
(Kobyliński and Przepiórkowski, 2008)	Random forest	Polish	E-Learning
(Westerhout, 2009)	Random forest	Dutch	E-Learning
(Fahmi and Bouma, 2006)	Maximum entropy	Dutch	Medicine
(Borg, 2007)	Genetic algorithms	English	E-Learning
(Storrer and Wellinghoff, 2006)	Rule-based	German	Technical texts
(Trigui et al., 2010)	Rule-based	Arabic	Question answering
(Przepiórkowski et al., 2007)	Rule-based	Slavic languages	E-Learning
(Miliaraki and Androutsopoulos, 2004)	Support vector machine	English	Question answering
(Saggion, 2004)	Rule-based	English	Question answering
(Walter and Pinkal, 2006)	Rule-based	German	Law

10

Towards a Supersemantic Analysis II - IntegreSSA

While Shallow SRL showed quite good results in the Propbank evaluation and the definition detection applications, there were also shortcomings to the approach: the clause splitting was based on a rule-based heuristic, which cannot be as powerful as a full-fledged syntactic analysis, likewise the detection of enumerations was only heuristic-based and prone to errors and finally the classification of the chunks was performed independent of each other making it possible to obtain argument combinations within a predicate-argument-structure that do not make sense (e.g. multiple ARG1s).

Most of these could be tackled by developing the heuristic used in Shallow SRL into a more powerful syntactic analysis tool or by replacing it by an existing one at the cost of sacrificing the efficiency of the system. Additional to these, however, there exist more fundamental problems with the general approach that was taken by Excerbt and adapted in Shallow SRL. While part of Shallow SRL's internal logic was extracted to a rule system, at the heart of it, it is still a machine learning system. As mentioned before, the problem with such systems is the fact that the models that they learn are encoded in a way incomprehensible for humans. Thus, they act as a black box and integration of expert knowledge and additional modules is tricky and at best indirectly possible.

Furthermore, a supervised classifier, like the ones used in Senna and Shallow SRL, always depends on available training data. For semantic role labeling, the Propbank corpus (Palmer et al., 2005) is the only available comprehensive corpus for training semantic role labeling [9]. Thus, all classifiers trained to role label are predetermined to follow the Propbank annotation scheme. One problem with this is the definition of large arguments mentioned before that makes syntactic post-processing necessary. Another, more fundamental one, is the fact that Propbank only creates PAS from verbs. Biological events, however, can also be described using adjectives or nominalizations (e.g. 'Foxp3-dependent activation of lrrk2'). These cases can never be captured when using a Propbank-trained classifier.

[9] Though it should be mentioned that with BioProp (Tsai et al., 2007) there exists a considerably smaller biomedical alternative that follows the same annotation guidelines as Propbank.

Likewise, Propbank does not account for nested events. Apart from that, it has to be pointed out that Propbank consists of annotations of financial news. Language might, however, differ between different domains. Consequently, a classifier trained on data from the financial domain might perform worse in the biomedical domain.

Based on all these considerations a second prototype was designed and developed. Because of the fundamental restrictions of the Excerbt approach, the underlying design was changed in this second prototype. Instead of using a supervised classifier a system efficiently combining a syntactic and semantic analysis was realized. The integration of syntactic and semantic information followed the approaches of feature structures, which define grammars including semantic features (see e.g. (Gardent and Kallmeyer, 2003; Kikui, 1992; Latreche, 2011)). This framework was designed to naturally integrate important NLP modules like word sense disambiguation, negation detection or anaphora resolution instead of having to apply them as mere post-processings. This way the syntactic analysis itself can benefit from semantic information provided by the modules. Moreover, this strong amalgamation of syntax and semantics is further promoted by including semantic information early on in the sentence analysis. In the spirit of this integration of modules and linguistic analyses, this second prototype is called IntegreSSA (integrated supersemantic analysis) .

10.1 Integrated analysis

For the work on IntegreSSA I collaborated with Felix Sappelt, a Master student I supervised, who implemented a German version of IntegreSSA for the analysis of German patient records (Sappelt, 2013). I designed the German sentence analysis together with Felix Sappelt and supervised the development process as well as the evaluation. This application of IntegreSSA will be presented in section 10.3. The technical realization of the system is based on a comprehensive linguistic analysis system that was provided by a Munich software company called Clueda[10]. Clueda's text analysis conforms with the ideas of a Supersemantic analysis. Based on this fundamental framework, I implemented different adaptations for the biomedical domain, additional PAS extraction levels and a biomedical event extraction system. The modular structure of the framework allowed the seamless integration of the anaphora resolution system described in chapter 6. In the remainder of this section, first the general concept of IntegreSSA and then the adaptations for the biomedical domain as well as the event extraction system are presented.

In order to guarantee a sound and comprehensive concept and framework the conception of IntegreSSA is based on a list of considerations that largely arose from the experiences made with Excerbt, Shallow SRL and the implemented tools described in the previous chapters. These considerations were reflected in several design decisions. In particular the following ones should be pointed out (cf. (Sappelt, 2013)):

- The artificial boundary between syntax and semantics was abolished. To this end the named entity recognition is preponed in contrast to Excerbt and many other text mining systems. This way lexical semantic information is available and can be beneficial during the sentence analysis. As mentioned before, his line of reasoning is adapted from feature structures. Furthermore, a semantic pattern recognition module was included at the beginning of the sentence analysis. This

[10] http://www.clueda.com

module tackles the detection of semantic concepts that naturally are not stored in a knowledge base, yet are still possible to detect without additional context information. An example of this are time designations like dates. In the scientific domain, such a pattern recognition module could e.g. be extended to detect citations. The pattern recognition module adds additional semantic information early in the analysis which can subsequently be used to make better decisions when analyzing the syntactic structure of a sentence. From a typical text mining perspective the pattern recognition module is a collection of NER algorithms like they are commonly used in many text mining systems.

- The system should be understandable as well as easily extendable and adaptable. Therefore, instead of having to deal with the blackbox character and training corpus dependency of machine learning systems, a rule-based approach was taken. This way events derived from nominalizations and adjectives could naturally be included without having to annotate training corpora. Furthermore, the adaptability of the system should be guaranteed by implementing IntegreSSA in a highly modular and flexible pipeline architecture. A so-called level stack allows to add and exchange modules with minimal effort. A module in this connection is a linguistic algorithm designed to solve a specific subproblem in the integrated analysis.

- Modules should be integrated at the stage were they make most sense. This includes e.g. modules like negations, word sense disambiguation, and contextual arguments like time or location information. Following this paradigm, the detection of time and location chunk is integrated early on in the pattern recognition stage. Likewise negation and WSD are integrated into the sentence analysis level stack instead of applying them as mere post-processings. Furthermore, word sense disambiguation is split into multiple tasks at different stages. Early on a comparison with POS tags is carried out. This way word senses with different POS tags like 'ache', the noun and 'to ache', the verb could be differentiated. At a later stage event information is taken into account to differentiate different word senses. For example, in a sentence 'ACHE is regulated in two neural cell lines by APP.' the information that the regulation event requires a gene as its theme (ARG1) is used to distinguish the gene 'ACHE' from the pain 'ache'. Thus, each relevant module to distinguish word senses is integrated where it would naturally fit into the processing pipeline.

- A way to solve multiple problems at once should be provided. The flexible level stack allows to solve problems at the stage that it makes most sense to solve them, e.g. by prepending the NER or by combining the extraction and contextualization of relations. In other tools like BioContext such contextualizations are performed in a post-processing step where the results of two individual tools are combined. In IntegreSSA, this is performed in one step when the relations between entities are extracted. As already pointed out in chapter 2, there are, however, interdependencies between different NLP tasks. In order to account for this, a backtracking mechanism to allow for down-stream effects of higher levels is included in the concept of IntegreSSA. This way e.g. chunks that look like enumerations could subsequently be split, if their components are needed as arguments in the event extraction step.

- The computational efficiency of the approach should still be guaranteed. Adding more syntactic analyses usually comes at the cost of higher processing times. Yet, for a large-scale application a certain degree of efficiency is required. For this reason, performance was constantly taken into

consideration during the implementation of IntegreSSA. This aspect was largely accomplished by the Clueda implementation that based the processing of their levels on specifically optimized grammars.

Based on these considerations, IntegreSSA was designed. A schematic overview of the developed prototype is given in Figure 10.1. As one can easily see, the design strongly followed the comprehensive supersemantic analysis network presented in Figure 2.4. While this prototype does not yet have all the features mentioned in Figure 10.1 (e.g. the learning capabilities and the section information are missing), the flexible architecture simplifies the integration of these modules.

In this framework, a text is analyzed by first disassembling it into its components. For this purpose, the sentences making up the text are extracted using a sentence detection module. The sentence, in turn, is split into the tokens it is made up of using a tokenizer. These tokens are then POS tagged in order to identify the part-of-speech of each token. All of these, sentence splitting, tokenization and POS tagging, are performed with the corresponding modules of the Java NLP analysis framework OpenNLP (The Apache Software Foundation, 2010). Based on these preprocessings the actual sentence analysis is conducted.

As mentioned before, the first step in this is the detection of named entities. In IntegreSSA, a dictionary-based NER approach was taken. The lexical information required for this was entered into and managed by the Clueda ontology technology, which guaranteed performant access and the possibility of extensively scaling up the resource in the future. The typical NER was accompanied by a pattern recognition step, hence the name NE/PR. Pattern recognizers already implemented in the Clueda implementation include date and time recognizers, a currency recognizer, and a percentage recognizer. The flexible implementation of the NE/PR module, however, allows to define and add additional recognizers with minimum effort. Two examples of these will be given in 10.3 where both the recognition of measurements and dosages were realized as recognizers in the NE/PR module.

The detected entities and patterns constitute the lowest level of chunks. A chunk is an accumulation of words that form a meaning-bearing whole. Chunking is a well-established linguistic procedure that has been widely applied to structure sentences (see e.g. (Grover and Tobin, 2006; Latreche, 2011; Ramshaw and Marcus, 1995; Sang and Buchholz, 2000)). In IntegreSSA, a chunk always has a type and potentially has meaning objects attached to it. Entities and patterns found by the NE/PR module form so-called 'KnownStuff' chunks. Each of these chunks is accompanied by a meaning object referencing the corresponding entry in the lexical resource. This way, the complete semantic information stored in the lexical resource is available during the whole sentence analysis process and can be used in the rules for building chunks and extracting contextualized relation. The common linguistic task of connecting entities in a text with a knowledge base is typically referred to as entity linking (see e.g. (Chen and Ji, 2011; Han et al., 2011; Zhang et al., 2010)). Exemplarily, for verbs frame set information is stored in the lexical resource that is essential for identifying the arguments of the verb. In such a frame set, the amount and the allowed types of arguments for a verb are stored. For the verb 'to activate' that can describe a positive regulation event e.g., the frame set contains the information that two arguments are needed. The frame set information for this was taken from Propbank (Kingsbury and Palmer, 2002).

In IntegreSSA, contextualized relations are extracted as predicate-argument-structures. By using the frame set information, the algorithm knows how many arguments it needs to look for. In the case of 'to activate' this is two (who activated what), while for 'to give' it would be three (who gave what to

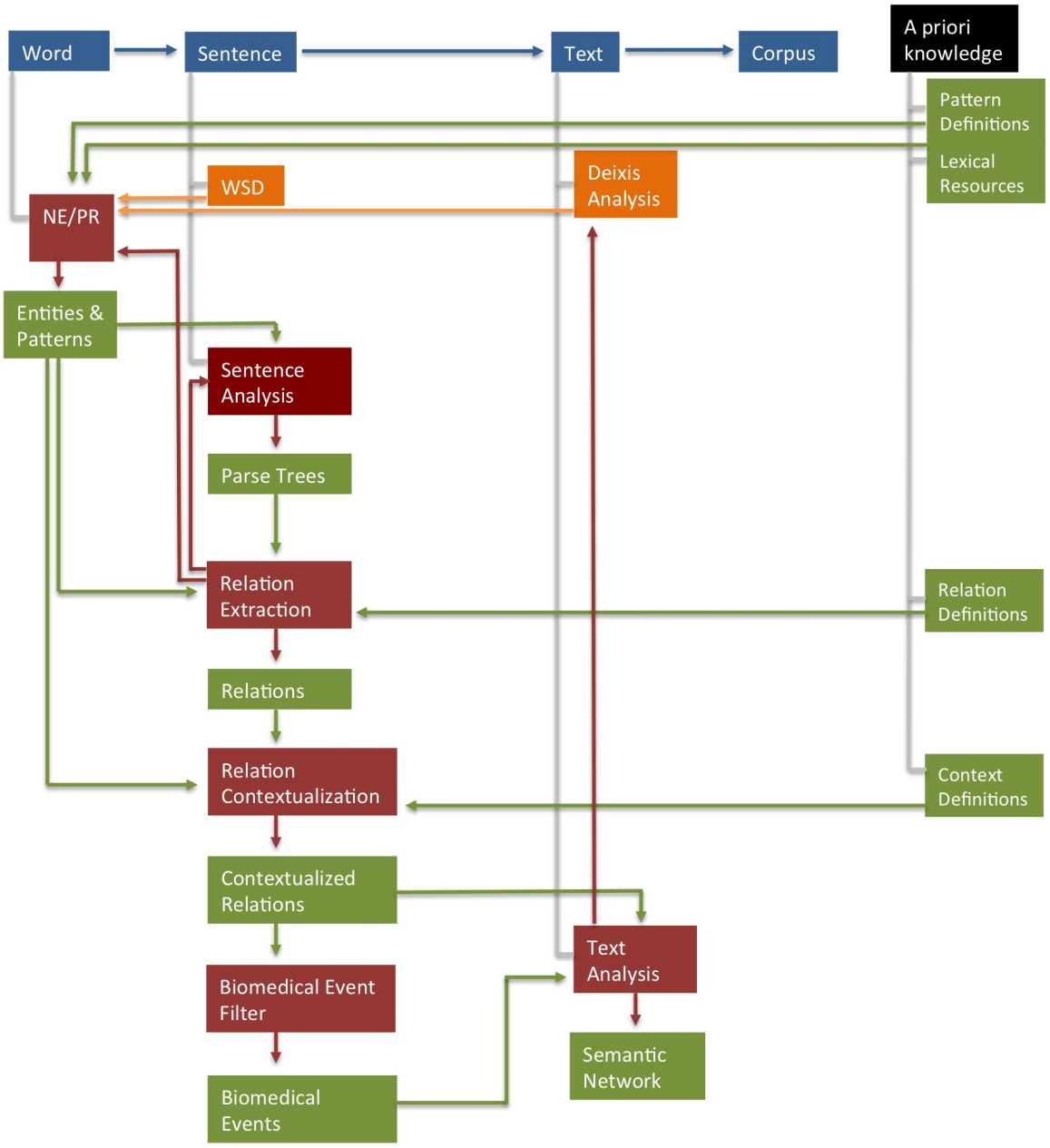


Figure 10.1: Schematic overview of IntegreSSA

whom) and for 'to die' only one (who died). Furthermore, the frame sets provide information about the types of the entities. This way an event PAS for 'to activate' is only produced if in Arg1 one of the required classes (protein or promoter) is found. If the same verb describes different events, multiple entries were included in the lexical resource resulting in multiple meanings that could potentially be appended to a chunk. This way the algorithm is presented with all possibilities and can rule out variants when they stop making sense in the course of the sentence analysis. Such decisions based on semantic constraints are in accordance with the semantical agreement of feature grammars (Latreche, 2011). In order to decide between multiple interpretations that all make sense, frequency information giving an a priori probability could be entered in the lexical resource and are considered during the event extraction step. In the biomedical case, this was done e.g. for the verb 'to express' which could be used to describe gene expression as well as transcription events.

The low level KnownStuff chunks and remaining tokens are then iteratively combined to more and more complex chunks. This Sentence Analysis module is realized by a modified context-free grammar (CFG) . CFGs were invented by Noam Chomsky in the late 1950s (Chomsky, 1959) and have been frequently applied ever since (e.g. (Charniak, 1997; Earley, 1983; Tomita, 1985)). Formally, a CFG is a 4-tuple consisting of an initial symbol S, a terminal vocabulary $\sum$, a non-terminal vocabulary V, and a finite set of rewriting rules R. The intitial symbol $S \in V$ and subsequently other non-terminal symbols from V can be rewritten by a series of rules defined in $R : V \rightarrow (V \cup \sum)*$. This way each non-terminal symbol is replaced by other terminal and non-terminal symbols until only terminals remain. In sentence analysis, such grammars are used to determine the syntactic structure of a sentence. Here, the terminal symbols correspond to words, the non-terminal symbols to phrases and the intitial symbol to the complete sentence. In IntegreSSA, a modified version of this CFG was used. Instead of terminal and non-terminal symbols actual objects of an object-oriented programming language were used in the implementation. Furthermore, the rewriting rules were adapted to work on functions of chunks instead of symbols. Here, arbitrary functions that map a chunk to truth values indicating whether or not it matches were allowed. These functions worked as a kind of algorithmic symbols which can be used to write rules similar to those in traditional CFGs. However, they are more powerful allowing to define constraints based on the tokens, POS tags, chunk types or even semantic properties of the chunks in question (cf. (Sappelt, 2013)). An example of how such a rule can be written is given in following:

$$< NC >::= \pi(c, DT)?\ \ \pi(c, JJ) *\ \ \pi(c, NN)$$

Here, a rewriting rule for a non-terminal noun chunk (NC) is shown. The function π is one of the algorithmic symbols mentioned above. It takes two parameters - a chunk and a POS tag - and matches if the given chunk possesses the given POS tag. In this case, an optional determiner (DT) is possibly followed by an arbitrary number of adjectives (JJ) and a noun (NN) . Using a variety of such rules different chunks and clauses are detected. The different chunks are detected with grammars organized in levels in the already mentioned level stack. For IntegreSSA, a total of 27 of such levels were created. The first of these levels is the NE/PR, the remaining ones are chunking or PAS extraction levels. The latter of which are also defined as grammars. Relations and their contextualizations (currently negation, time and location) are extracted in the form of PAS at different stages of the analysis pipeline. This allows to use the semantic interpretations of the PAS in the remaining sentence analysis in the same way the lexical semantics of ontology entities and patterns from the NE/PR step could be used. An example of the results of such a sentence analysis is given in Figure 10.2.

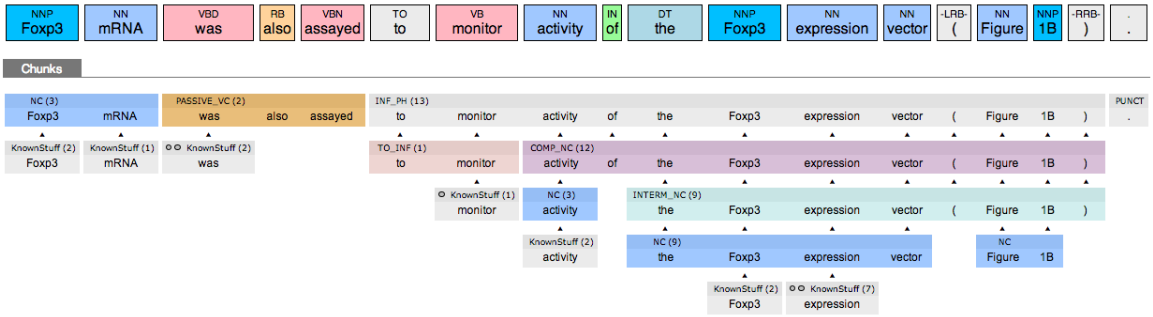


Figure 10.2: Example output of the IntegreSSA sentence analysis.

Finally, in IntegreSSA the PAS are filtered for those corresponding to biological events. To determine which PAS are relevant the training set of the BioNLP corpus was examined for verbs, nouns and adjectives that can trigger biological events. For each of these, the corresponding frame set information was manually included in the ontology. In total 293 frame sets were created and categorized in the eleven different event categories of BioNLP. The different categories were: Binding, Gene Expression, Negative Regulation, Positive Regulation, Localization, Phosphorylation, Protein Catabolism, Protein Modification, Regulation, Transcription and Ubiquitination. A filter for biological events was written that accepted a PAS only if it matched the corresponding frame set information of an event predicate.

Both the filtered biological events and the normal PAS derived from the single sentences are then used in the text analysis module. Here, the different relations are combined to form a semantic network representing the knowledge of the underlying text. This network is an abstraction of the original text. Semantic networks are a typical representation of semantic relationships between concepts that has been widely applied in different linguistic and AI applications (see e.g. (Havasi et al., 2007; Niemann et al., 1990; Shapiro and Rapaport, 1986; Sussna, 1993)). In IntegreSSA, this network is represented as a topic map. Such an abstract representation is in line with van Dijk's idea of information reduction which he discussed in the context of his macrostructure framework for discourse analysis (van Dijk, 1977). Van Dijk proposed different macrostructure operations to reduce semantic complexity. For example, he proposed the deletion of rather irrelevant attributes of arguments. In an utterance 'a little town', for instance, van Dijk proposed to omit the attribute 'little' in order to simplify the macrostructure of the text. In accordance with this, the semantic network created based on topic maps avoids attributes that are not explicitely entered in the ontology. The inclusion of further paradigms from pragmatics and discourse theory in the supersemantic analysis is planned for the future. Exemplarily, the different illocutionary speech acts of Austin described in section 2.2 result in different sentence types like questions, commands or normal statements. A detection and distinction of these is planned. This would on the one hand support the sentence analysis and could on the other hand be used in the text analysis in order to use the statements derived from these appropriately. In accordance with this, a distinction of different discourse types could be added in a similar fashion.

The general text analysis framework described above was extended and modified in several ways to meet the requirements of the biomedical event extraction task IntegreSSA was designed for. An overview of these extensions and modifications is given in the following:

Tokenization modifications

The tokens used in IntegreSSA are slightly more fine-grained than the original ones. First of all, the OpenNLP tokenizer used by the Clueda framework sometimes leaves brackets attached to words. This is corrected in InegreSSA. Furthermore, a certain type of chunk is separated into three different chunks due to the adjective event extraction discussed further below. In BioNLP expressions like the following are considered biological events:

CD40-induced upregulation of CD80

In this example two nested events should be detected. The first one is the upregulation of CD80 and the second one is the induction of this upregulation that was triggered by CD40. Thus, the word 'CD40-induced' includes both the predicate and the Arg0 of the predicate-argument structure describing the second of these events. Since the PAS detection mechanism implemented in IntegreSSA aims at identifying and labeling chunks for the different roles in a PAS, it is necessary to split this word up. This way the different parts become token chunks that can be identified as predicate respectively Arg0 of a PAS. For this reason, such formulations that contained relevant predicate trigger words were split into three tokens: the noun, the hyphen and the adjective.

POS tagging modifications

Like most NLP tools the POS tagger used in IntegreSSA (OpenNLP POS tagger) was trained on financial news texts. This typically produces problems when the tagger is used in other domains with words it never encountered during its training phase. For this reason, in order to improve the quality of the POS tags a series of post-processings was performed. This way the tagger was adapted to the biomedical domain.

Furthermore, POS taggers highly depend on the capitalization of words. Unfortunately, in scientific articles headlines are often capitalized, which frequently leads to missclassified tokens. Such a behavior is especially problematic for verbs. Many capitalized verbs are tagged as nouns and thus can be used to derive meaningful biomedical events from them. In order to correct for this, another post-processing step is required. In total, in IntegreSSA the following post-processings of POS tags were conducted:

- POS tags of event words were changed to their corresponding correct POS as determined by the ontology entry of the word, whenever the assigned POS tag clashed with the ontology information and the form as it was given in the text.
- POS tags of proteins as given in BioNLP were corrected to proper nouns.
- POS tags of words containing digits as well as letters were changed to proper nouns. In addition to the BioNLP proteins other proteins and compounds were commonly misclassified by the original POS tagger. This post-processing counteracted this.
- The bracket tokens detached in the tokenization modifications were POS tagged with the corresponding -LRB- and -RRB- tags.

Biomedical ontology

Since the approach taken in IntegreSSA depends on the early availability of lexical semantic information, the creation of a biological ontology was required. In contrast to the ontology used in Excerbt, the IntegreSSA ontology is hierarchical. This hierarchy was made use of at different stages. For example, in the anaphora resolution module described in chapter 6 the parent-child information were used to detect appropriate antecedents. This way, for example, 'Eomesodermin' could be matched to all its parents in the hierarchy like 'T-box', 'transcription factor' and 'protein', while other proteins like 'il-10' would also be matched to the term 'protein', however not to 'T-box' but instead to its own parent nodes like 'Interleukin'.

The created biomedical ontology consisted of 69 classes and 5076 individuals. Classes are all umbrella terms that can describe different individuals. They are the possible anaphors in the anaphora resolution algorithm. Individuals are the leafs of the ontology tree. They correspond to real entities in the world and actual descriptions of activities. Thus, the type of biological event (e.g. a gene expression event) was categorized as a class, while the actual term describing the event (e.g. the verb to express) was categorized as an individual. The largest amount of individuals are so-called unclassified verbs. They were provided by Clueda and consisted of verbs and their frame set information. These entries were used in the course of the sentence analysis and PAS extraction. Additional to these, in the course of this work frame set information for verbs describing the events covered in the BioNLP competition were included in the ontology. These consisted of 293 entries.

BioNLP entity recognizer

Since task 1 of the BioNLP competition is a pure event recognition task that is largely independent of named entity recognition, the relevant proteins for which events should be extracted are given in all texts. In order to integrate this information in the IntegreSSA workflow, an additional protein recognizer was implemented and added to the NE/PR module. The recognizer reads the protein information given in the BioNLP input files and accordingly creates KnownStuff chunks with the corresponding protein meanings. This way the correct proteins are detected and the semantic information that this chunk refers to a protein can be used throughout the analysis.

Chunking modifications

Since Clueda did not provide me with all required functionalities, it became necessary to add certain chunking and PAS extraction levels in order to be able to capture all relevant information. Concerning the chunking the following modifications and additions were implemented:

- In expressions like 'CD40-induced upregulation of CD80', adjectives are built by combining a noun and an adjective with a hyphen. Like mentioned before both the adjective and the noun need to be detected since they can form the cause and the predicate in a biomedical event. Thus, the tokenization separates them to make them detectable individually. However, the different tokens would break the usual sentence analysis because the knowledge that the construct noun-hyphen-

adjective corresponds to an adjective is missing. Consequently, to account for this, an additional AdjectiveHyphenLevel was implemented that builds adjective chunks from such utterances.

- An ellipsis chunk was added that detects utterances of the form ', but not ...,'. As could be seen in the small sample evaluation performed in the evaluation of the Negatome text-mining approach in section 4.5, such ellipses could make up as much as 10% of negated events. Furthermore, detecting these chunks prevents the elements of it to be mistaken for arguments in other PAS and thus increases the performance of the sentence analysis.
- A domain-specific phenomenon of scientific articles is the use of certain forms of citations. In many of these the cited element is put in brackets in the middle or at the end of a sentence but is not part of the sentence in a linguistic sense. Thus, in order to analyze the sentence correctly these citations have to be detected as such. In the articles in the BioNLP data numbers in squared brackets are used to indicate citations. In order to detect these a level building chunks from digits surrounded by these brackets was implemented. Furthermore, to properly detect biological entites like the virus '[HTLV-I]' as nouns, all other expressions in such brackets were combined to noun chunks.
- Since Clueda did not provide me with a comprehensive enumeration detection for noun chunks, a heuristic similar to the one in Shallow SRL was implemented. Noun chunks before the first and after the last verb that look like an enumeration were combined to noun chunks. Furthermore, such patterns in between two verbs were combined based on an educated guess. If the last noun chunk was immediately followed by a verb, it was considered likely that there are two clauses combined with and 'and' and thus no enumeration was detected. Furthermore, in order to indicate that multiple PAS need to be build from arguments containing an enumeration, an enumeration meaning was constructed and attached to the resulting chunk, which referenced the components of the enumeration.

Additional PAS extraction levels

Analogously, for PAS extraction certain levels needed to be implemented, so the most common ways, in which biomedical events were described in the BioNLP data, were covered. For this purpose, the following PAS extraction levels were added:

- While nominalizations derived from genitive constructions like 'the activation of Foxp3' could be detected by the sentence analysis pipeline, nominalizations without a genitive structure like 'Foxp3 activation' could not be detected. To counteract this shortcoming an additional nominalization PAS extraction level was implemented that extracted PAS from noun chunks if they contained a protein and a fitting biomedical event predicate.
- A PAS extraction level for adjective predicates was implemented. Here, noun chunks containing adjective predicates were analyzed. Both normal adjectives and those with the noun-hyphen-adjective construction mentioned above are used for analysis. This way an utterance like 'CD40-induced upregulation of CD80' results in a PAS with the predicate 'induced', Arg0 'CD-40' and Arg1 'upregulation of CD80', the last of which in turn contains an additional PAS with 'upregulation' as a predicate and 'CD80' as Arg1.

- Apart from within noun chunks adjectives can also express events in other forms. Two examples of these are given in the following two utterances:

  “LMP1 was detectable in LMP1 transgenic B cells, …”
  ‘Deletion of the FKH domain, critical for nuclear localization, …”

  In the first case, the adjective appears as an object of a ’to be’-verb. The subject of the verb should be the Arg1 in this case. The PC belonging to the verb should be the ArgM-Loc. The second case is similar to the first. However, here the ’to be’-verb is omitted and the adjective appears within the resulting ellipsis. Again the NC in the beginning, which would be the subject of the ’to be’-verb should be the Arg1 of the PAS generated from the adjective. In order to cover both of these cases, a second adjective PAS extraction level was implemented that detects these utterances.

- A level extracting PAS from the ellipsis chunks mentioned before was implemented. In an expression like ’Overexpression of Foxp3, but not of deltaFKH, …’ this level creates a PAS containing of ’Overexpression’ as predicate and ’deltaFKH’ as Arg1.

Apart from these additions, there were slight modifications made to the format of the predicate-argument-structures in order to match the representation used in BioNLP. For example, in the original Clueda sentence analysis pipeline PAS were defined on chunks which lead to predicates like ’have been activated’. In contrast to this, BioNLP referenced only the head word of the verb chunk (’activated’). In such cases, the format was adapted to match BioNLP in order to provide a meaningful evaluation.

Furthermore, as mentioned above, for PAS containing arguments with enumeration meanings for each of the referenced elements one PAS containing the referenced element as the corresponding argument was created.

Nested events

Another functionality, not yet, provided by the Clueda pipeline was the extraction of nested events. In order to circumvent this short coming, a mechanism to simulate the nesting of events was implemented. Whenever biological events were detected, a meanig representing this event was attached to the corresponding predicate. This way the semantic information about the event could be used when checking whether this event would fit as an argument for the superordinate event.

10.2 Evaluation

IntegreSSA was evaluated and compared to Excerbt on various levels and for various quality measures. First, the quality and the processing speed of the two sentence analysis modules were compared. For this purpose, an evaluation on Propbank was conducted. Both systems were tested on the same Clueda testing system in order to allow comparability. The results of the processing speed in relation to sentence length can be seen in Figures 10.3 and 10.4. With an average processing time of 40 and 41 ms respectively, IntegreSSA and Senna showed very similar performance. Looking at the time vs. sentence length plot one can, however, make out an exponential increase of processing times with increasing sentence length for Senna. This could turn out to be problematic when using the tool in a

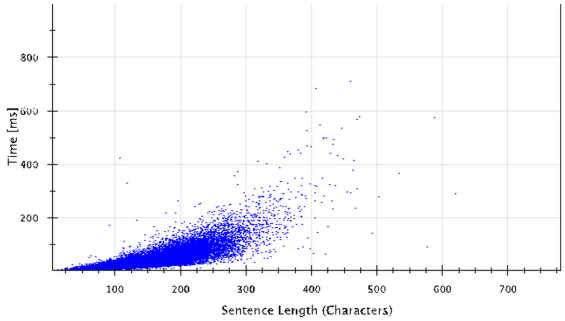


Figure 10.3: Processing times of Senna in relation to the length of a sentence.

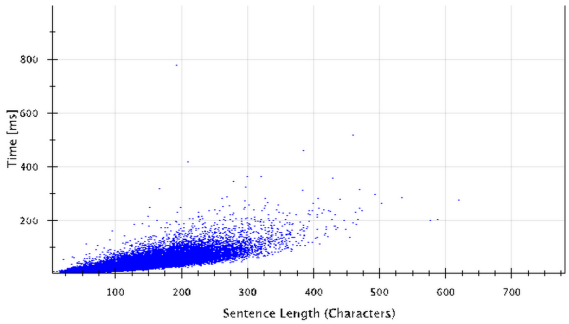


Figure 10.4: Processing times of the IntegreSSA sentence analysis in relation to the length of a sentence.

scientific domains with possibly more complicated and complex sentences. IntegreSSA, on the other hand, showed a more linear increase of processing times with respect to the sentence length.

The quality comparison of Senna and the sentence analysis module of IntegreSSA was conducted on the full corpus of Propbank. Since IntegreSSA thus far focusses on the extraction of obligatory arguments, the evaluation was concentrated on the extraction of predicates, Arg0 and Arg1. Here, Senna detected 95.9% of the predicates of all PAS correctly. For Arg0, a precision of 0.66, a recall of 0.99 and a F-measure of 0.79 was reached. For Arg1, a precision of 0.69, a recall of 0.99 and a F-measure of 0.81 was reached. In comparison to this, the sentence analysis module of IntegreSSA detected 82.8% of all PAS correctly. For Arg0, the precision, recall and F-measture values were 0.53, 0.69 and 0.60 respectively. For Arg 1 these were 0.42, 0.78 and 0.55.

For two reasons, these values should, however, be treated with caution. First of all, Senna has an advantage because the data for testing mostly consisted of its own training data. The reported results obtained when the training data of Senna is removed from the testing corpus is a F-measure of 0.75 for all arguments (including predicates). It should be noted that this still includes a - eventhough smaller - advantage for Senna, since both training and testing set are based on the same domain (Wall Street journal corpus) while the rules of IntegreSSA were not developed based on Wall Street journal texts. Secondly, the annotation scheme of Propbank differes from the predicate-argument-structure definitions used in IntegreSSA. As mentioned before, Propbank has a tendency to define large

arguments. For example in the following sentence, the Arg0 encompasses two appositions, a relative clause and a prepositional phrase:

> {Los Angeles, California, at the West Coast of the United States, which is famously known for Hollywood, the home of the entertainment industry,}$_{Arg0}$ {has to deal}$_{Predicate}$ {with a major debt problem}$_{Arg1}$.

IntegreSSA, on the other hand, would only detect 'Los Angeles' as Arg0 and consider all the other information as additional information concerning 'Los Angeles' but not as argument of 'has to deal'. As mentioned before, this extensive argument definition caused a problem when using Propbank-trained tools for event extraction systems and thus motivated the inclusion of argument restriction rules in Excerbt implemented by Robert Strache (Strache, 2012). In order to soften the effect of these different definitions the quality values mentioned above are derived using a soft matching approach that considers two arguments to be the same if at least 80% of its tokens overlap. While this attenuates the effect, in cases like the one above the more sensible Arg0 'Los Angeles' derived by IntegreSSA would still not be considered correct. To my knowledge there do not exist comprehensive alternatives to Propbank. For this reason, the event extraction evaluation presented in the following might be more meaningful for assessing the quality of the two sentence analysis systems compared with the Propbank performance evaluation.

The event detection of IntegreSSA was evaluated on the BioNLP 2013 shared task 1 data (Nédellec et al., 2013). BioNLP is a series of shared tasks that was initiated in 2009 and ever since held every two years. BioNLP sees itself as a "community-wide effort to address fine-grained, structural information extraction from biomedical literature" (Nédellec et al., 2013). The shared task is typically accompanied by a workshop in which the obtained results are presented and the different experiences are shared. In 2013 this workshop was held in Sofia. The BioNLP competition always consists of different tasks. Teams are able to participate in as many of these as they like. In the original 2009 shared task there was one task focussing on event extraction offered. This was extended to five different tasks in 2011 and six in 2013. The different tasks of 2013 were the following:

- [GE] Genia Event Extraction for NFκB knowledge base
- [CG] Cancer Genetics
- [PC] Pathway Curation
- [GRO] Corpus Annotation with Gene Regulatory Ontology
- [GRN] Gene Regulation Network in Bacteria
- [BB] Bacteria Biotopes

All of these tasks were event extraction tasks from different domains. Text with a marked set of biomedical entities was always given as input and events containing these (and potentially additional ones) needed to be extracted. Besides the domain, the tasks vary in their focus. The GRO task focusses on the construction of a complex semantic ontology, the GRN task additionally requires to build gene regulation networks, and the BB task largely focusses on location contextualization. Furthermore, the GE task additionally includes an anaphora resolution and a negation/speculation detection subtask, which the participating teams could additionally solve.

For the evaluation presented in the following, the core event extraction annotations from the development data of the GE task were used. While the anaphora resolution data was used for the evaluation of the anaphora resolution system presented in chapter 6, it was left out in the event extraction evaluation presented in the following. This was done in order to guarantee comparability with the Excerbt results that are used as a baseline (since Excerbt lacks an anaphora resolution mechanism). The core event extraction metric was the major metric used within the BioNLP competition. To extract a core event the predicate, the type of the event and the primary arguments (referred to as causes and themes) need to be correctly identified. Secondary arguments refer to additional contextual information like the site at which a protein modification event took place. These were considered in a second task called event enrichment which is not the focus of the evaluation presented in the following.

Precision, recall and F-measure were calculated for the evaluation. The results are shown in Table 10.1. In addition to the core event extraction results, the values for detecting the correct event type at the correct position (but possibly making errors at the arguments) are shown. As can be seen, Excerbt scores rather low F-measure values of 0.07. The poor quality is largely due to the extremely low recall. This is in accord with the shortcomings discussed in chapters 4 and 9.

Table 10.1: Performance of IntegreSSA in comparison to Excerbt on BioNLP shared task GE task data.

System	Prec. ev. type	Rec. ev. type	F-m. ev. type	Prec. core	Rec. core	F-m. core
IntegreSSA	0.67	0.35	0.46	0.56	0.29	0.38
Excerbt	0.51	0.04	0.08	0.45	0.04	0.07

IntegreSSA on the other hand, reaches between five and six times the quality values of Excerbt (F-measure) both for core event extraction and for detecting the correct event type at the correct position. For both tasks the precision values are nearly twice as high as the recall values. This can be partially explained by the rule-based approach and the annotation peculiarities of BioNLP discussed later in this chapter.

Looking more into detail one can see that Excerbt performs reasonably at detecting causes in the few cases where events were detected (precision of 0.45, recall of 0.46 and f-measure of 0.45), but often fails at detecting themes (precision of 0.1, recall of 0.1 and f-measure of 0.1). In contrast to this, IntegreSSA performs more evenly with a performance advantage for detecting themes reaching precision/recall/f-measures of 0.26/0.25/0.25 for causes and 0.46/0.44/0.45 for themes.

Likewise, the event type evaluation shows clear advantages for IntegreSSA over Excerbt. IntegreSSA even increases its large lead to 0.38 F-measure points compared to the core event extraction. This measure was considered because it might serve as an indicator for the suitability of the system for using it in a semi-automatic fashion like as an assistant tool for manual annotators (like in chapter 4). In such a set-up, it is important to detect the correct events at the correct places in order to speed up the annotation process. Smaller mistakes in the arguments of the events are less problematic since they could be quickly fixed manually. Such scenarios seem to be the most practical use case of temporary text mining systems considering that none of the currently existing systems (see Table 10.4) could exceed a precision 0.63 and thus could be considered reliable enough for a completely automatical use.

In comparison to the participants of the BioNLP contest, IntegreSSA scores in the lower midfield. Considering that IntegreSSA is a prototype built in few months and only containing very few of the supersemantic modules proposed in this thesis as well as not yet any learning capabilities these results seem rather satisfactory. For comparison, the top two entries in BioNLP have been developed since at least four years, the third placed BioSEM for at least one. Since BioSEM, like IntegreSSA thus far, is rule-based, this might be seen as an indicator that rule-based systems could be a strong competitor for the established machine learning approaches or at least that a larger feature base derived from a more sophisticated rule system can be very beneficiary for all kinds of text mining systems.

In addition to the quality assessment of IntegreSSA, an error analysis was performed in order to identify the reasons for remaining missed or missclassified events. This analysis revealed that there does not seem to be one predominat reason for the errors but instead that it is a variety of many small error sources which often could be fixed in the scope of a rule-based system with additional effort. An excerpt of identified error sources is given in the following:

- Prepositional predicates were not yet included in IntegreSSA. Beside events from verbs, nominalizations, adjectives and ellipses that were included in IntegreSSA, some events in the BioNLP data set are derived from prepositional clauses. Take for example the following sentence:

  > RT-qPCR was used to determine gene expression levels of il-6 and cxcl8 in response to PMA following inhibition of NF-kappa.

  In such formulations, the preposition is usually the predicate, the theme is an event described by a clause or a nominalization and the cause is the noun chunk of the PC of the preposition which also often describes an event in the form of a nominalizaton. In the given example, both the 'following' and the 'in response to' are such prepositional predicates. The 'in response to' is an expression that acts like a preposition. Such prepositional events were not yet included in IntegreSSA and were thus a source of missed events.

- For effective communication, all parts of an expression that could be infered from the context could be omitted. These omissions are known as ellipses. For ellipses that leave out verbs, an event extraction module was implemented in IntegreSSA. In addition to this, however, there exist ellipses that omit an argument of the event. In the following sentence, e.g., the theme of the binding event was omitted and needs to be inferred from the context:

  > Deletion of the carboxyl-terminal forkhead (FKH) domain, critical for nuclear localization and DNA-binding activity, abrogated the ability of Foxp3 to suppress NF-kappaB activity in HEK 293T cells, but not in Jurkat or primary human CD4+ T cells.

  Here, a localization and a binding event of Foxp3 should be detected. The theme (Foxp3) of the predicates that are given in nominalizations (localization, binding) is omitted in the apposition and can only be inferred from its occurrence later in the sentence. These special cases of ellipses are not yet implemented in IntegreSSA.

- Events can be given as parts of definitions. Such definitions can occur in the form of clauses, often with predicates derived from 'to be' or expressions like 'act as', or in the form of appositions. The following two utterances give examples of these two cases:

> Stat6, the IL-4 target, ...
>
> LMP1 acts as a constitutive signal through ligand-independent oligomerization ...

In the BioNLP data, the first expression constitutes a regulation event, the second a binding event with 'oligomerization' as predicate and 'LMP1' as theme. Such formulations are not yet implemented in IntegreSSA.

- For comparison reasons, the anaphora resolution module (see chapter 6) was not used for the evaluation. Including a comprehensive anaphora resolution system would improve the resutls significantly.
- In addition to these factors, an integrated system always depends on the quality of its components. Since IntegreSSA promotes low-level analyses, this problem is reduced. However, there still exists errors in sentence splitting, tokenization and POS tagging, for which external tools were used.

In addition to this, the quality values of all systems suffer from certain inconsistencies or peculiarities of the annotation of the BioNLP data. Some of these that were encountered during the error analysis should be discussed in the following:

- In event extraction systems, the event type is often closely linked to the predicate. In Excerbt e.g., the event type was fully defined by the verb that occurred as predicate. IntegreSSA went a step further and used the complete predicate-argument-signature for its event definition. This way a verb like 'express' could describe a gene expression event if the theme is a protein and a transcription event if the theme is a mRNA. In BioNLP, however, there are different event types defined on the same predicate-argument-signatures. For example, the predicate 'overexpression' with a protein as theme is defined 17 times as positive regulation and 5 times as gene expression event in the train data set and 11 times as gene expression and 2 times as positive regulation in the development data set. While there might exist hints in the context of the expressions which of these interpretations should be used in a given situation, the fact that the distinguishing factor is not annotated within BioNLP makes it harder to deal with these kind of situations. Moreover, the same formulations with the same predicate-argument-structure are inconsistently annotated as events at all. Take for example the following two consecutive sentences:

  > Deficiency of RUNX1 or RUNX3 resulted in markedly reduced TGF-beta - mediated induction of FOXP3 mRNA in naive CD4+ T cells compared with control cells transfected with scrambled siRNA. The level of FOXP3 mRNA was further reduced when both RUNX1 and RUNX3 were knocked down in naive CD4+ T cells during their differentiation to iT reg cells (Fig. 1 B).

  In the second sentence, 'FOXP3 mRNA' is considered a transcription event. In the first sentence, however, the same expression 'FOXP3 mRNA' was not annotated. It might be interpreted that the second 'FOXP3 mRNA' is a metynonym that actually refers to the transcription of FOXP3 and not the product of this transcription, the mRNA. In such a case, however, a comprehensive annotation should indicate this (possibly in a similar fashion as anaphoras are resolved). Including such complicated constructions without also annotating the underlying linguistic concepts further complicates the event extraction.

- Sometimes the annotation deviates from a linguistic semantic analysis. Take for example the following sentence:

  > Phosporylation of Ser127 on NF-kappaB by PKA recruits the transcription co-activator, p300.

  Here, BioNLP defined a positive regulation with the predicate 'by' and a nested phosphorylation event without a cause as theme. Linguistically, however, one would consider the 'by PKA' as an agent (thus in this connection a cause) of the phosphorylation. Such deviations from the linguistic norm make it more complicated to build systems following the BioNLP annotation scheme.

- While anaphoras of nouns are annotated within the BioNLP data, deitic verbs are not. Instead of leaving such terms out for the event extraction, however, pro-verbs like 'do' are used as predicates in events. An example of this can be seen in the following sentence:

  > Wild-type NleH1 expression significantly reduced TNF-induced RPS3 S209 phosphorylation, whereas the K159A mutant failed to do so (Fig. 7b).

  In the BioNLP annotation, a negative regulation containing the predicate 'do' is expected. From my perspective, it seems more sensible to use 'reduced' as a predicate for both events (the one including the wild-type and the one including the mutant) and include an anaphora with the trigger 'do' and the antecedent 'reduced' in the annotation.

The analysis of annotation peculiarities in BioNLP revealed another advantage of rule-based systems. When developing a rule-based system one seems to be more inclined to study the annotation scheme in more detail. Since contradictory annotations make rule-creation impossible, one can often quicker identify such annotation shortcomings.

10.3 Patient Record Analysis

In addition to the English version of IntegreSSA, a German version was implemented by Felix Sappelt under my supervision in the course of his Master thesis (Sappelt, 2013). Both versions use the same technical framework but differ with respect to the used models for tokenization and POS tagging as well as the language-dependent chunking levels and NE/PR recognizers. This prototype (subsequently referred to as German IntegreSSA) was applied to the problems of measurement and named entity recognition in German patient records.

As already pointed out in chapter 5, text mining systems can be used as a backbone for diagnosis support systems. In such systems, patient files are analyzed semantically in order to derive a comprehensive patient profile. This profile can then be directly compared to known profiles of diseases using weighted vector similarity measures or can be used as input for machine learning algorithms. Using such decision support systems can help a physician in his diagnosis of patients. Such systems become especially useful in cases were diseases are very rare or very complex. In the first case, the physician simply lacks the experience about the disease in question, since he might only encounter it once in his whole career. In the second case, the interplay of many factors might hint towards the disease. Because of

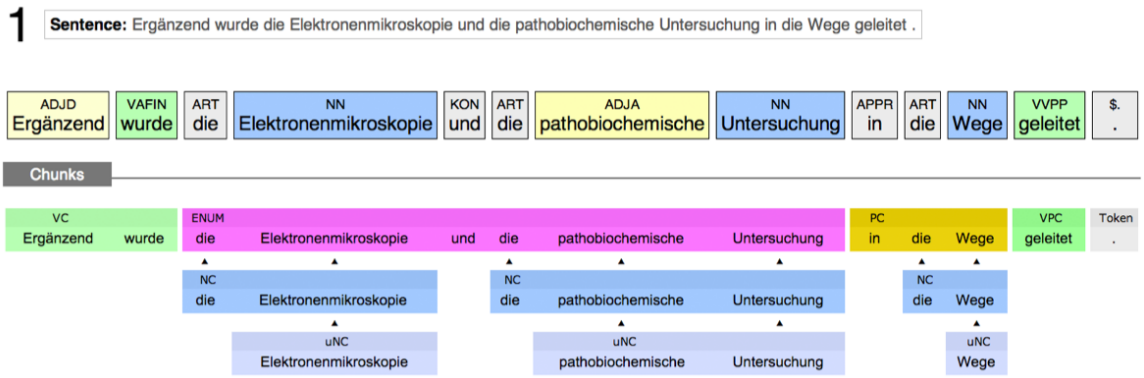


Figure 10.5: Example output of the German IntegreSSA chunking analysis.

this multitude of factors, however, it becomes difficult to recognize the specific patterns. In such cases, an automated system might help to avoid overlooking diseases.

In chapter 5, only nominal attributes (which symptomes are present) were considered for the disease profiles. A lot of important information, however, is given in the form of numeric values. Among the most prominent of these are measurements and dosages of all forms. In order to make these accessible to the analysis of a decision support system, in German IntegreSSA a special measurement detection NE/PR module was implemented. In addition to that, the foundation of a full-fledged analysis of German patient records was created by setting up a German biomedical ontology and implementing German sentence analysis chunking levels. For this purpose, the following additions and changes to the original IntegreSSA were made:

Table 10.2: Units supported by the German IntegreSSA measurement recognizer.

Measurement	Units
Length	Metre, Inch
Mass	Gram, Ounce, Dalton
Substance amount	Mole
Electrical current	Ampere
Temperature	Kelvin, Celsius, Fahrenheit
Volume	Cubic metre, Litre
Voltage	Volt
Energy	Joule
Information	Byte
Reactive amount	International Unit

- A biomedical ontology was created consisting of 7245 drugs or substances and 199 measurands. The drugs and substances were derived from the Rote Liste [11]. Including synonyms 17515 terms for drugs and substances are findable. The measurand entries were manually created by Felix Sappelt. Including synonyms a total of 405 measurand terms are findable.

[11] http://www.rote-liste.de

- The OpenNLP tokenizer and POS tagger were used with the corresponding German models. It should be pointed out that the POS tagger model was trained on the Tiger corpus, which uses the STTS POS tag set. This German tag set with its 53 different tags is considerably more powerful than its English pendent, the Penn Treebank tag set, with its 36 different tags, which was used for the training of the English version. An overview of the STSS POS tag set is given in Appendix F.
- A measurement recognizer was implemented. This recognizer detected combinations of numeric values or ranges and units. A step-wise rule-based approach was implemented to detect first the numeric values and ranges (using a regular expression) and the units (using a context-free grammar) and then to combine these to measurements. An overview over the covered units is given in Table 10.2. Each of these units could be combined with any SI prefix or its symbol ranging from Pico (p) to Exa (E).
- Each detected measurment was accompanied by a special Quantity object. A Quantity object is a Meaning object that automatically internally transforms the measurement into the appropriate SI unit. Thus, both 254cm and 100" would be converted to 2.54 meters. This allows to easily perform calculations on the detected measurements.
- German chunking levels were implemented. Since PAS extraction is not yet included in the German IntegreSSA, this implementation consists only of a smaller level stack of eight levels. An overview of the used chunk types is given in Table 10.3. An example of the chunking results is given in Figure 10.5.

Table 10.3: Chunk types used by German IntegreSSA.

Chunk Type	Symbol in Visualization
Noun Chunk	NC
Prepositional Chunk	PC
Verb Chunk	VC
Adjective Chunk	AdjC
Adverb Chunk	AdvC
Known Stuff Chunk	KnownStuff
Enumeration Chunk	Enum
Relative Clause Chunk	RelativeClause
Subordinate Clause Chunk	SubordinateClause
Verb Particle Chunk	VPC
Parentheses	Parens
Undetermined Noun Chunk	uNC
Conjunction Chunk	Conjunction
Relative Clause Start	RCStart
Subclause Body and End	SCEnd
Token Chunk	Token

For the development and the evaluation of German IntegreSSA, a set of 82402 anonymized patient records was provided by the Friedrich Baur Institut[12]. In most of the cases, the patients described in these records were treated for neuro-muscular diseases. From this set, a random sample of 100 records containing 4857 sentences was selected. The sentences were filtered for whether they contained at least one digit. 267 of these sentence were found to contain actual lab values. These sentences were split into two separate test sets (in the following referred to as S1 and S2) to hint towards the variance of the quality assessment. Both the NE/PR modules and the chunking were evaluated on these sets. For the ontology-based NER and the measurement recognition accurracy values were calculated. The evaluation of the chunking was done exemplarily on a collection of randomly chosen samples due to the lack of a gold standard.

For the measurand recognition, the evaluation revealed an accurracy of 60.6% on S1 and 65.2% on S2. Manual evaluation of the results showed that lacking measurands or synonyms in the ontology were the main reason that prevented higher values. Adding these entries increased the accurracy to 99.1% and 97.8% respectively. The remaining missed entities could be attributed to mistakes in sentence splitting and tokenization. For the drug and substances recognition, accurracy values of 62.7% on S1 and 50.8% on S2 were measured. The missed instances could be attributed to missing drugs in the Rote Liste ontology in about half the cases. In the second half of these cases, the missing entities could be manually associated with at least one ontology entry. Here, the format of the Rote Liste posed a considerable problem. The entries are designed for human understanding and not for an automatic recognition, which lead to entries like the following:

> Pravasin protect 10 mg/-20 mg/-40 mg

Such formulations, however, are never written in this form in patient records by physicians. For this reason, an ontology creation algorithm was applied when deriving the ontology from the Rote Liste, which produced synonyms like ’Pravasin protect’ and ’Pravasin protect 10 mg’ from the above entry. Yet, the names used in the patient records were sometimes even further altered, which accounted for about half of the missed entities. In the case above, e.g., ’Pravasin’ was used by a physician, which was not created as a synonym by the ontology creation algorithm. In order to capture all relevant synonyms considerable domain knowledge would be required.

For the measurement recognition, accurracy values of 89.5% on S1 and 88.7% on S2 were obtained. Additionally, 5.7% and 7.7% respectively of the measurements were partially detected. The error analysis of the missed measurements showed that unusual formats (using a ’.’ instead of the correct ’,’ as a decimal separator or having fractional units with a left out numerator as in ’70/min’) and missing units (e.g. ’mmHg’ or ’nmol/UCS’) were the main reasons for the remaining errors.

The sample analysis of the chunking showed many correctly formed chunks ranging from rather simple NCs to more complex enumerations and subordinate clauses and even for not grammatically well formed sentences. Still, there existed a large range of rarer situations that result in chunking errors. Examples of these are prepositional phrases with an adjective as head word like ’nach oben’, detached verb particles like in ’gab an’, and the correct attachment of postposed auxiliary verbs like in ’eingenommen habe’. In addition to that, the chunker was naturally error-prone when it was fed with erroneous tokens or POS tags, and like any sentence analysis had problems with very complicated

[12] http://www.baur-institut.de/

linguistic problems like PP-attachment . Examples of the chunking results can be found in Figure 10.5 and Appendix G.

The patient record analysis is an example of how the flexible architecture of IntegreSSA can be easily extended to include additional analysis modules for additional tasks and even different languages. Furthermore, it shows that the kind of rule-based analyses promoted in IntegreSSA can provide very promising results on real data.

10.4 Related Work

IntegreSSA is intended to be the prototype of a supersemantic analysis system. Therefore, all of the approaches mentioned to be related to supersemantics in section 2.6 are also related to IntegreSSA. Furthermore, since IntegreSSA includes a semantic role labeling module also all approaches mentioned in section 9.5 should be taken into consideration as related to IntegreSSA. Additional to these the participants of the BioNLP shared task 2013 are listed in Table 10.4 to provide further information about the scientific background of IntegreSSA.

Table 10.4: Results of the official BioNLP shared task GE task (Nédellec et al., 2013) submissions for core event extraction.

Team name	Reference	Approach	Recall	Precision	F-Measure
EVEX	(Hakala et al., 2013)	SVM-based pipeline	45.44	58.03	50.97
TEES-2.1	(Björne and Salakoski, 2013)	SVM-based pipeline	46.17	56.32	50.74
BioSEM	(Bui et al., 2013)	Rule-based pipeline	42.47	62.83	50.68
NCBI	(Liu et al., 2013a)	Joint pattern matching	40.53	61.72	48.93
DlutNLP	(Li et al., 2013)	SVM-based pipeline	40.81	57.00	47.56
HDS4NLP	(Liu et al., 2013b)	Joint SVM	37.11	51.19	43.03
NICTANLM	(MacKinlay et al., 2013)	Joint pattern matching	36.99	50.68	42.77
USheff	(Roller and Stevenson, 2013)	Mixed pipeline	31.69	63.28	42.23
UZH	(Schneider et al., 2013)	Rule-based pipeline	27.57	51.33	35.87
HCMUS	(Pham et al., 2013)	Mixed pipeline	36.23	33.80	34.98

11

Discussion

11.1 Integrated systems

One of the main ideas promoted in this thesis is that a comprehensive linguistic system requires an integrated system that combines all relevant syntactic, semantic and pragmatic analyses of linguistic subtasks in a way that each of these analyses benefits from the other ones. Only such an integration enables an accurat interpretation of statements in the corresponding context in which they were uttered. This idea fits in with different current developments in adjacent fields like machine learning and bioinformatics, which should be mentioned here in order to provide already implemented examples of how sensible integration can increase the performance and flexibility of classification problems.

In machine learning, deep learning and multi-task learning were two trends that delivered very promising results parallel to my work on this thesis. Deep learning refers to a way of combining machine learning models usually in subsequent layers. Here, each layer solves its own supervised or unsupervised learning problem. Often, the first layers are unsupervised learning layers that learn a better representation of the input data. Using this representation the subsequent layers then solve a classical supervised classification problem. The unsupervised stage could be realized by autoencoders or Restricted Boltzmann Machines (both special kinds of artificial neural nets), while the supervised stage could be realized by a support vector machine or a belief network. But the main idea behind deep learning does not depend on the actual choice of models but instead on the combination of learning tasks (Bengio et al., 2013).

For this reason deep learning is also closely related to multi-task learning. Multi-task learning refers to the approach of learning a model or part of a model for multiple different problems at once. The idea behind it is to reach a problem-independent abstract internal represenation that is useful for solving different problems of the same field. This way one can “exploit commonalities between different

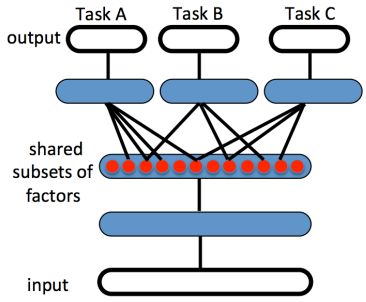


Figure 11.1: Prototypical architecture of a multi-task learning system that learns representative factors for each subtask and thereby strengthens the generalization abilities of the model. Figure taken from (Bengio et al., 2013).

learning tasks in order to share statistical strength, and transfer knowledge across tasks" (Bengio et al., 2013). Multi-task learning has already been shown to improve the performance in different natural language processing tasks (e.g. (Carlson et al., 2010b; Collobert et al., 2011a)). It can naturally be combined with deep learning by training the first unsupervised layers as general representations for multiple tasks and then stack different supervised learners on top of this for each of the individual learning problems (Bengio et al., 2013). See Figure 11.1 for an illustration of such a set-up.

Between 2010 and 2014, deep learning algorithms became the state-of-the-art machine learning algorithms. They outperformed other approaches like random forests and support vector machines in several competitions (e.g. (Ciresan et al., 2011; Roux et al., 2013; Stallkamp et al., 2011)) and also industrial leaders shifted their focus in this direction (e.g. Facebook hired the deep learning expert Le Cun to head its new artificial intelligence lab (Metz, 2013) and the Google Brain project is based on deep learning algorithms (Le et al., 2012; Markoff, 2012)). The main focus of deep learning application was on image and speech recognition (see e.g. (Dahl et al., 2012; Le et al., 2012; LeCun et al., 1989)) thus far, but there have also already been transfers to text analysis problems. Interestingly, here Senna is the best known example. In this connection it should be pointed out that the before uttered criticism of Senna-based Excerbt should not be understood as a criticism of Senna. Senna's ability to detect the semantic roles of elements of sentences as defined in Propbank without the time-consuming use of rich syntactical features is - to my knowledge - umatched by any role labeling system. The criticism of Excerbt instead focussed on the point that for event extraction Senna on its own does not suffice but additional syntactic analysis would be required, which was not sufficiently implemented in Excerbt. Senna accomplishes its good results by using a combination of multi-task and deep learning as mentioned above. The different tasks learned simultaneously are part-of-speech tagging, chunking, named entity recognition, and semantic role labeling (Collobert et al., 2011b).

With an increase of NLP tasks, the need to organize the different modules sensibly becomes more important. For the implementation of the concepts presented in this thesis and the integration of the here evaluated modules, a comprehensive framework with clearly defined interfaces that enables productive collaboration in larger teams would be beneficial. This brings us to the second current trend in machine learning and bioinformatics: the use of comprehensive multi-purpose workbenches. Such workbenches comprise a variety of processing algorithms that make up modules, which a user can easily combine to powerful processing pipelines. The well-defined interfaces and the graphical user interface greatly simplify the modification of processing pipelines and the reuse of existing modules.

In machine learning, Weka (Hall et al., 2009), Rapidminer (Rapid-I, 2014) and dotplot (dotplot, 2014) are among the available workbenches. Furthermore, a more general data processing workbench that is gaining popularity in the field of bioinformatics is KNIME (KNIME development team, 2013) (see Appendix B for an example of a KNIME workflow). The creation of such a workbench for a supersemantic analysis seems very beneficial.

However, existing technologies suffer from two shortcomings that make them impractical for supersemantic analyses: a very rudimentary format concerning the interfaces and a commitment to sequential processing. When it comes to the format for passing information from one module to another, the workbenches favor maximally general formats. For example, in KNIME the interface format is always a table. This has the advantage of being able to use these modules for arbitrary data, but might cause problems for complex structures like the ones required in linguistic processing. Using a more specialized application programming interface (API) would allow to formulate more complex and more linguistic specific data structures in a more convenient way. Semantic information (like nested PAS) e.g. seems to be most naturally represented in complex hypergraphs. As intermediate formats in the process of sentence processing, additionally alternative interpretations of subgraphs coupled with conditions or probabilities under which they hold might be required. While all of these could be translated back into tables, the overhead for doing so and the large amount of conventions that would have to be applied and followed seems to make this representation very impractical. Since information representation is a very central topic of complex linguistic analysis, the following section contains more thoughts on such adequate representations.

Secondly, the commitment to sequential processing poses a problem for using existing workbenches for linguistic analysis frameworks. As already pointed out in section 2.5, there exists both upwards and downwards causation between the different levels of linguistic structure and meaning. This necessitates that a supersemantic framework provides functionality to solve multiple problems with various interdependencies simultaneously. The deep/multi-task learning approaches provide one way of realizing this. Others might be optimization techniques with a variety of constraints or iterative approaches with correction properties like the Backtracking level stack proposed in IntegreSSA. However, a purely sequential processing pipeline like provided by most current workbenches does not meet these prerequisites. While some of these workbenches might be diverted to simulate the correction ability, the other two possibilites seem more complicated. Possibly, the machine learning workbenches adapt to deep learning algorithms because of their great popularity at the moment. However, if these adaptations come with the required degree of flexibility remains questionable. Furthermore, in a comprehensive supersemantic workbench, one would like to exchange the mechanism of simultaneous processing just in the same way one exchanges a machine learning model in current workbenches. For this additional work in the field of information science is required to provide a unified representation that can be used for each of the processing mechanisms. How such a representation could look like is discussed in the following section. Once such a representation is found, workflow management of simultaneous workflows or workflows with correction loops become possible.

This can lead to a powerful linguistic workbench that allows flexible integration and exchange of interdependent modules in the future. Among the current comprehensive natural language processing toolkits LingPipe (Carpenter and Baldwin, 2011), nltk (Bird, 2006), OpenNLP (The Apache Software Foundation, 2010) and Stanford NLP (University, 2011) should be mentioned. None of these, however, is a workbench or provides the possibility to model such interdependencies at the moment.

11.2 Knowledge Representation

The most fundamental problem of sentence analysis is ambiguity. Different approaches exist to tackle different instances of this ambiguity. Syntax tries to apply general formal rules of how to form grammatical sentences in order to determine the structure of a sentence, semantics and logic try to test the resulting interpretations for coherence in the context of existing knowledge, and statistical methods try to guess the best interpretation based on frequencies of observed patterns of language use. Each of these approaches proved itself to be successful for certain tasks and each of them should be part of a comprehensive sentence analysis system. To exemplify this further, take a look at the interpretation of the following sentence:

> "The boy saw the man with the binoculars."

This sentence is one of the typical sentences used in linguistics to exemplify the problem of PP-attachment. Here, the prepositional phrase 'with the binoculars' could either be attached to the verb meaning that the man is using the binoculars or the object meaning that the boy is in possession of them. Besides this syntactic ambiguity, however, there is also a lot of lexical ambiguity: The term 'the boy' could refer to a boy mentioned earlier or the planned first movie of a serial killer trilogy produced by Elijah Wood, the verb 'saw' could be the past tense of 'to see' or the infinitive of 'to saw', and 'the man' could refer to a man mentioned earlier, the gene symbol of a the mandarin gene or to the slang term for an authority. Furthermore, there could be a new term introduced in this sentence. 'The man with the binoculars' might be a new band, movie or artist that the man went to see.

Now each of the before mentioned studies is able to solve certain of these ambiguities or at least help to find the most sensible interpretation. Syntax delivers the basic structure associating part-of-speeches and a hierarchical structure with the elements of the sentence. Furthermore, it can detect that 'saw' must be the past tense of 'to see' because otherwise there would be a mismatch concerning the subject that is in singular and the verb form. However, it cannot distinguish between the different interpretation where the binoculars belong to the man or the boy or even where 'the man with the binoculars' is a named entity. Semantics can rule out that 'the boy' refers to the movie since movies do not see things, but in a strict sense of semantics it cannot distinguish whether the gene, authority or man were seen by the boy. Logic, however, can rule out the version that the binoculars belong to the gene and additionally that the gene was seen using the binoculars because genes are too small to be seen using binoculars. Thus, using such an inference logic can deduce that the expression 'man' could not belong to the gene. Finally, statistical methods can be used to further support the analysis of the sentence. Using association patterns one could determine that 'man' more often refers to an actual man than to the authority and possibly that prepositional phrases at the end of the sentence more often belong to the verb than the object. Using these information one could come up with a most likely interpretation of the sentence.

As should have been illustrated by this example, all four analysis practices - syntax, semantics, logic and statistics - play an important role in the analysis of a sentence. Yet, there is a lack of systems combining methods from all four fields. One major reason for this is probably the required effort to integrate the existing tools for the different fields. One of the difficulties one would be faced with in such a case would be to find a comprehensive data structure that is able to represent all the required

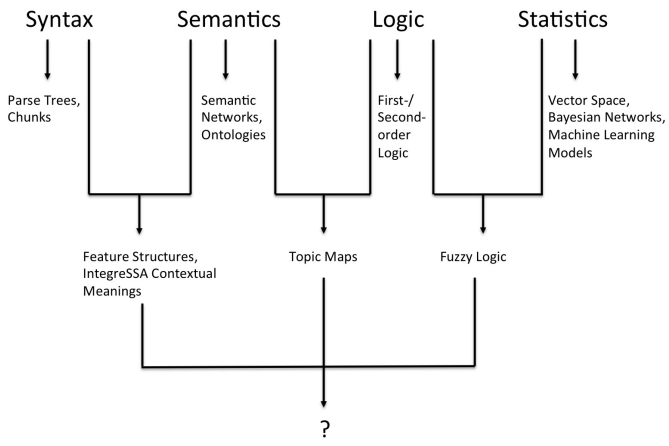


Figure 11.2: Overview of different approaches to knowledge reprentation.

information. This endeavour is especially complicated since all of these approaches use different systems of knowledge representation (see Figure 11.2).

Syntax analyses typically use parse trees or chunks to represent the structure of a sentence, semantics often uses ontologies or semantic networks to represent the interdependencies of concepts, logic most commonly uses expressions formulated in first- or second-order logic and statistics uses vector spaces (see chapter 7) or machine learning models or in order to represent the probabilities derived from the statistical analyses uses Bayesian networks. Finding a representation that subsumes multiple of these knowledge systems is not trivial. There are several representations that bridge two of these fields. For example, Fuzzy Logic combines logic with uncertainties (that could be derived with statistical methods), Topic Maps combine at least certain constraints with semantic networks if scopes are used and Feature Structures enrich syntactical analyses with additional sometimes semantic features. In addition to that the contextual meanings (a semantic meaning object together with its instantiating chunk) of IntegreSSA is a data structure combining syntax and semantics. A data structure combining all four of them, however, is - to my knowledge - missing. Thus, finding an appropriate data structure remains an important goal of future work and could also be seen as a prerequisite of the before mentioned hypothetical supersemantic workbench because working with inappropriate knowledge representations inevitably complicates and restricts the use of this knowledge as could be seen with the missing syntactical information in the topic maps used as input for the anaphora resolution system in chapter 6 where syntactical information was missing and accordingly a feature depending on them could not be implemented.

Figure 11.3 provides a sketch how such a data structure could look like. Here, the main representation of a graph was chosen. The syntactic interpretations derived from the sentence are shown in green, the semantic ones in blue. Furthermore, the events derived from the sentence are shown in orange. The box containing the X, Y and Z is a semantic frame set that describes how the verb to see can be

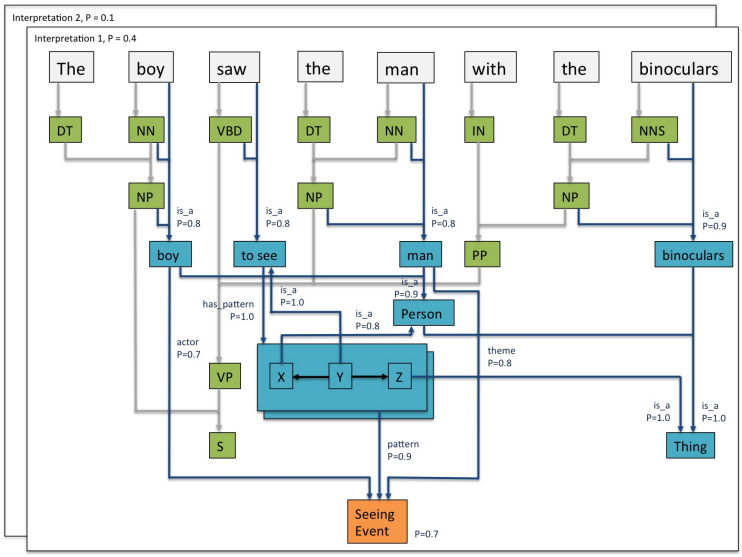


Figure 11.3: Sketch of a possible data structure integrating syntactic, semantic, logical and statistical information.

used in a sentence. There is another box hidden behind it indicating that one verb can have multiple of these frame sets. Different possibilities to interpret the sentence are represented by explicitely creating alternative interpretations and assigning probabilities to them. Thus at each junction where a decision on how to combine tokens has to be taken, all possibilities are evaluated and weighted with their corresponding probabilities. Depending on the program such a data structure would be used in, low probability alternatives could be pruned or all alternatives could be played through until the end. Probabilities were chosen to serve as representations both for the statistical and the logical interpretation modules. Therefore, all semantic edges in the graph are assigned probabilities. If they violate a logical constraint these probabilities would be set to zero or to allow for errors in the logic module to very low non-zero values. For this purpose, edges or more general patterns could be transformed into logical propositions, which then in turn could be checked by a typical logical solver. For example, the is_a relation between man and person translates to "$\forall x.Man(x) \Rightarrow Person(x)$". Based on the probabilities of the edges the probabilities of the events and of the whole interpretation are derived. The data structure is indepent on the actual mechanism to calculate the probabilities. An intuitive way to calculate them would be frequency counts on annotated corpora like Penn Treebank (Taylor et al., 2003) or Propbank (Kingsbury and Palmer, 2002), which provide syntactic and semantic information for large corpora. Focussing on rather general representations like graphs and probabilities would make the data structure independent from the processing modules it is used with. Every semantic or syntactic processing module would simply need to produce weighted edges and in case of ambiguities multiple interpretations. Every logic or statistics module would simply need to derive probabilities to modify these probabilities based on the analysis of subgraphs.

A second issue concerning the representation of knowledge is how to represent the meaning of terms. In dictionaries, terms are usually just listed without any interpretation of their meaning or possibly associated with a broad superclass that associates the terms with a label like 'protein' or 'disease'. Ontologies are more complex allowing for a hierarchy of terms that allows to make more fine-grained

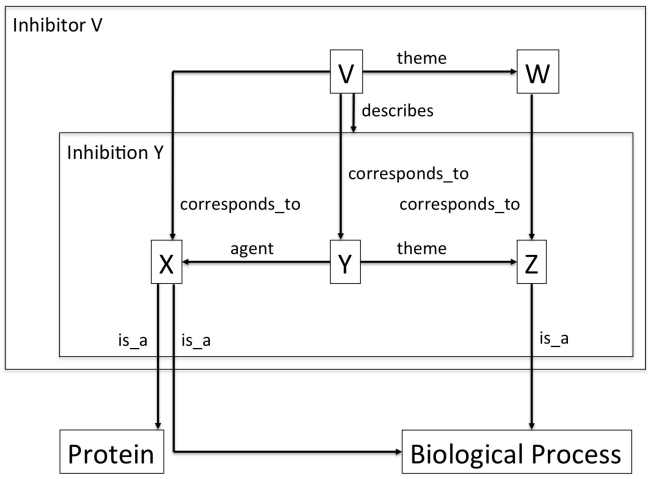


Figure 11.4: Sketch of a possible graph pattern to represent the meaning of the term 'inhibitor'.

distinctions. For most terms, however, there exist different ways of categorizing them. For example, a disease could be categorized as rare or frequent depending on its prevalence, as genetic or non-genetic depending on its cause or as lethal or non-lethal depending on its consequences. For this reason, certain ontologies also allow multiple hierarchies turning the representation into an acyclic graph. While such an organization already covers a lot of information, it still seems to fail at representing certain term meanings. Take for example a word like 'inhibitor'. An inhibition usually consists of two players: the one inhibiting and the one being inhibited. The description of an inhibition typically consists of three terms: the two players of the inhibition and the term describing that it is an inhibition (e.g. 'p53 inhibits lrrk2'). The term inhibitor, however, subsumes both the descriptive term of the inhibition and the player causing the inhibition. This situation is similar to the hyphen combined adjectives described in sections 10.1 (e.g. in 'p53-dependent'). While in the hyphen case it was possible to separate the two elements again, with a term like 'inhibitor' this is not possible. Instead one would need to create a way that allows for terms to be represented in a more powerful way.

A natural way of doing this that is convenient in the context of a largely graph-based representation might be graph patterns. Such a pattern would be a graph just like the semantic network derived from a text analysis. This graph pattern, however, is underspecified meaning that some of its nodes are not entities of the real world but rather variables about which certain features are known. For example, the term inhibitor could be represented by the pattern shown in Figure 11.4.

Here, inhibitor is a pattern VW, where W is the theme of V and V describes the inhibition pattern. The V corresponds to both the X and the Y in the inhibition pattern. Since X is either a protein or a biological process also V has to match one of those classes. Likewise, since Z is a biological process so is W. The Y in the inhibition pattern could further be defined as a negated activation pattern. This way all knowledge could be set into relation to each other up until very basic physical and linguistic entities. For clarity the probability values on the edges are left out. The patterns could be extended to represent

time, location or logical constraints in order to widen the range of representable knowledge. Such a data structure could provide the foundation of a comprehensive automatic reasoning framework because it integrates information even if it is uttered in different forms.

11.3 Analysis efficiency

The efficiency of linguistic analysis is an important factor for the evaluation of a system. Many text mining systems develop in a direction that adds more and more extraction modules extending processing times. Exemplarily, BioContext that added multiple contextual analyses needed half a year for processing MEDLINE (Gerner et al., 2012). Such immense processing times make it very complicated to keep ones system up-to-date. However, especially new information is of particular interest to scientists and thus the extraction of this is an important application field for text mining systems. Even more, such long processing times make improvements employing reprocessings impractical.

In this work, the efficiency of analyses was always a deciding factor in choosing underlying tools and in some cases required the own implementation of rather low-level tools in order to provide adequte processing times. This approach worked well for the systems developed in the course of this work. With the extension of the framework to further analysis modules like proposed by the integration of logic and statistics (Section 11.1), efficiency, however, becomes an issue again. When ambiguity is explicitly embraced by managing all possible interpretations simultaneously like proposed in Figure 11.3, the processing time of the analysis scales with the number of alternatives (depending on the stage at which the ambiguities occur). In the spirit that existing modules should always benefit from additionally added ones, in this section I propose a way of how statistical modules might improve the processing speed.

A text consists of different types of information. In this section a distinction between explicit and implicit information is introduced to explain this phenomenon. While facts or speculations are explicitly mentioned, often there is more information given implicitly. Such implicit information can be hidden ”between the lines”, it can be information left out that can be infered from the context (van Dijk calls such implicit propositions MISSING LINKS (van Dijk, 1977)) or can be associations based on the distribution of utterances within a text or corpus. In particular, the association patterns mentioned in Section 7.1 are a form of this implicit information. The graph patterns introduced in Secton 11.2 are also a way of representing these association patterns when they are paired with a probability value that indicates the strength of the association. Text passages can be transformed into such patterns by generalization. Here, the terms in the sentence could be replaced by their parent terms (the ones that they are connected to via a is_a relation). Each combination of words and parent terms within the pattern could be created and for each the occurrence frequency over a large corpus could be estimated. Take, for example, for the following three sentences:

> Lrrk2 activates p53.
> P53 activates lrrk2.
> P53 activates man.

Here, patterns for *activate(Protein, Protein)*, *activate(lrrk2, Protein)*, *activate(p53, Protein)*, *activate(p53,Person)*, *activate(Protein, p53)*, *activate(Protein, lrrk2)*, and *activate(Protein,man)* would be

created (using the typical notation of predicate logic where the verb in front of the brackets describes the predicate and the terms between the brackets describe the objects). Since 'man' is ambigue the pattern was created for both interpretations of 'man'. If there would be a reliable method to resolve this, the *activate(p53,Person)* would of course be left out. However, since only the most frequent patterns are important it should not influence the procedure too much if all interpretations are kept in. All of these created patterns would have the frequency of 0.33, only the pattern *activate(P53,Protein)* would have 0.66 and *activate(Protein,Protein)* would have 1.0. Thus, generalization of the players in this case provided a protein activation pattern like it would also be used in biomedical event extraction systems. Of course one would have to constrain the generalization by penalizing a too high level of generalization to avoid a dominance of patterns like *activate(Thing, Thing)*. The frequencies of the patterns could be interpreted as a form of familiarity of an expression and could be used in the form of probabilities in the data structure of Figure 11.3.

Furthermore, these probabilities could be used to speed up the processing. Instead of always evaluating all possibilities only the most likely one could be further analyzed until the end or until it becomes less likely than one of its alternatives. Thus, in such a procedure the association patterns guide the processing order and a dynamic pruning of unlikely interpretations. Such an approach is a form of lazy evaluation, a programming technique used for more efficient algorithms. The principle here is to evaluate an expression only when and if it is needed. Furthermore, this approach is similar to the efficient best-first search. In this search, a graph is explored by always expanding the most promising node as defined by a heuristic evaluation function. In the case of the association patterns guided analysis, this evaluation function would be given by the probability values of the association patterns.

In addition to that, there is evidence that such an approach also resembles the way humans process language. An intuitive example of these are garden path sentences. In such sentences, the reader typically first interprets the meaning of the sentence in one way until at the end of the sentence a reinterpretation becomes necessary. Garden path sentences are used by psycholinguists to illustrate that humans normally process sentences sequentially. Some examples of garden path sentences are the following [13]:

> The horse raced past the barn fell.
> The old man the boat.
> The government plans to raise taxes were defeated.

In the first sentence, the horse is first though to actively race past the barn. This interpretation is changed to a passive 'being raced past the barn' when the 'fell' required a subject. In the second sentence, 'old' is first interpreted as an adjective and 'man' as a noun, which then needs to be changed to a noun and a verb respectively. Correspondingly, in the third sentence the 'plans' are first interpreted as a verb and later as a noun. Thus, in case of garden path sentences humans also first process the most likely interpretation and then move back to an initially less likely one if the first one became incoherent.

A second phenomenon hinting towards the effects of associations in language understanding is priming. Priming is an implict memory effect that lets people quicker understand certain utterances if they are preceded by certain other utterances. For example, people are quicker at whether the String 'NURSE' is

[13] Examples taken from http://en.wikipedia.org/wiki/Garden_path_sentence

a well-formed word if it is preceded by the word 'DOCTOR' (Meyer and Schvaneveldt, 1971; Meyer et al., 1975; Schvaneveldt and Meyer, 1973). Priming can occur with a wide range of associations. Associative priming works on terms that co-occur with each other like cats and dogs in the expression 'it is raining cats and dogs' (Matsukawa et al., 2005). Thus, priming works strongly on idioms. But it also works on a semantic level, e.g. between dog and wolf (Reisberg, 2005). By extracting probabilities of meaningful patterns such a priming effect is transfered to a text mining system as well and thereby improve and speed up the analysis.

The topic of including statistics in linguistic analyses was famously discussed by Chomsky and Norvig. At the Brains, Minds, and Machines symposium during the MIT's 150th birthday party [14], Chomsky was asked in a panel discussion about his opinion of the increasing use of statistical methods like machine learning in the field of linguistics. Chomsky answered that there "have been some successes, but a lot of failures" (Chomsky and Pinker, 2011) and went on critizing that the models do not help in understanding the underlying principles and are therefore irrelevant for science. Furthermore, he argued that humans do not base their analysis of texts on probabilities but rather semantic and syntactic rules (Chomsky and Pinker, 2011). On his website, Peter Norvig countered Chomsky's criticism and stated that by analyzing the properties of statistical models also scientifically meaningful insights could be gained. Norvig went on to claim that language is a stochastic phenomenon and that therefore probabilistic models would be an obvious choice. Chomsky's critic of existing stochastic models was illegitimate according to Norvig because the models are too simplistic in order to explain all of language (Norvig, 2011). Instead, " [w]hat is needed is a probabilistic model that covers words, trees, semantics, context, discourse, etc." (Norvig, 2011)

Interestingly, Norvig's vision of a wholistic language learning system as stated here fairly resembles the ideas of a supersemantic analysis put forward in sections 2.5, 11.1, and 11.2 of this thesis. Consequently, I disagree with Chomsky, who states that in contrast to statistical learners, humans do not use frequency-based associations when understanding language. Intuitive associations and plausibility stem from the occurrence frequencies of the involved elements. Examples like priming or the fact that native speakers of a language do not need to think about the grammatical rules of this language point towards the explanation that strong associations can bypass parts of a rule-based analysis. However, Chomsky's criticism of the blackbox character of statistical models has to be emphasized. This was one of the reasons that during the course of the work on my thesis I more and more moved away from statistical models (like in the WSD tool and Shallow SRL) to more expressive often largely rule-based systems (like in IntegreSSA). Additionally, the dependency on annotated gold standards that often poses a practical problem should be pointed to in this connection. In contrast to Chomsky, however, I considered this as only a first step in order to gather all relevant information that are then in turn needed for a comprehensive integration with statistical models.

11.4 Implications of literature-based science

Text mining is a literature-based science and as such its opportunities are closely linked to the properties of the available literature. In this section, two of these properties will be discussed in more detail. The first implication of literature-based science is the fact that the amount of knowledge extracted by text

[14] http://mit150.mit.edu/symposia/brains-minds-machines

mining is bound by the amount of available literature. Secondly, the reliability of text mining results is depended on the quality of the literature as well.

It is a fact that there exists a considerable gap between published scientific literature and scientific literature that is freely available for text mining. In chapter 4, in the sample evaluation of the Negatome 1.0 results the ratio of existing and available information was approximated as 5 to 1. Another indicator of the proportion of available biomedical literature could be the ratio of the free section of Pubmed (Pubmed, 2014) to PMC (PMC, 2014). Here, the size of Pubmed serves as an approximation of all articles and the size of the free section of PMC as an approximation of freely available full-texts. With over 23 million Pubmed citations and around three million PMC citations, this ratio amounts roughly to 8 to 1. Independent of which of these two approximations is closer to reality, the extent of this problem becomes obvious. The wast majority of published biomedical knowledge is unreachable for non-commercial text mining systems.

The problem of this availability gap is noticed by many people in the text mining community. To name just a few examples, John Wilbanks argued in Nature in his comment “License restrictions: A fool’s errand” against the Creative Commons attribution license (Wilbanks, 2013), likewise Michael W. Carroll warned in his article “Why Full Open Access Matters” about publishers trying to commercialize text mining (Carroll, 2011) and Murray-Rust et al. reviewed how “researchers and information technologist[s] are blocked by legal and contractual barrie[r]s” (Murray-Rust et al., 2012) and submitted their results to the UK’s Hargreaves report on intellectual property reform.

The success and progress in turning these efforts into legislation, however, is varying. The Hargreaves report (Hargreaves and Office, 2011) demanded copyright exceptions among others for the purpose of text mining. This report was broadly accepted by the British government (Osborne et al., 2011) and is currently implemented into legislation. In contrast to this, progress on a European level was faltering. The issue of problematic copyright laws was addressed in a European Commission’s initiative called “Licenses for Europe”. In this structured stakeholder dialogue one working group was specifically concerned with text and data mining. However, the “discussions f[e]ll apart” (Van Noorden, 2013) and the initiative ended with five of the involved citizen organizations stating that they were “compelled to conclude that 10 months of meetings have largely failed to identify any solutions which can be backed by all, or even the majority of, stakeholders involved. It is evident that there is very little consensus among stakeholders about the appropriate approach to making EU copyright law and practice fit for the digital age. It is unclear as to how licensing solutions can provide a significant improvement to a copyright system that has been widely recognised as being inefficient and out of date” (Centrum Cyfrowe et al., 2013). The failure of the “Licenses for Europe” initiative shows that the path to more text mining friendly copyright laws is still long and weary and that a lot of effort of scientists is still required to convice the involved stakeholders of the necessity of this endeavour.

While the solution to the availability gap is largely outside of the range of influence of the involved researchers, the second implication of literature-based science, the quality dependency, is not. Linguistic analyses typically assume that the text they are analyzing is grammatically well-formed and truthful. In practice this assumption is, however, often violated. Typographical errors can occur everywhere where texts are written and in a domain like science where many authors are non-native speakers grammatical errors are fairly common. Luckily, such variations from well-formed utterances can be tackled by text mining approaches themselves. Many approaches to text normalization have been

proposed that correct these errors (e.g. (Castellanos, 2004; Nahm, 2004; Nahm et al., 2002)). Especially, with the increasing use of social media posts and short messages for text mining analyses such methods move more and more into the center of attention (e.g. (Beaufort et al., 2010; Kaufmann and Kalita, 2010)). For these cases, text mining can correct the errors by itself and furthermore by using text mining in the form of grammar- and spell-checkers, text mining can improve the quality of future additions to its literature basis as well.

While the progress concerning syntactic errors is already significant, the correction of semantical errors is more complicated. Science deals with the unknown and is therefore prone to false interpretations or the publication of flawed experimental results. In the introduction of this thesis, this was exemplified with the case of the conflicting results about the role of β amyloid in inclusion body myositis patients (Greenberg, 2009). With an increasing sophistication of text mining methods, however, also this phenomenon might be counteracted. The most pomising approach to this is automatic reasoning (already introduced in section 7.1). By performing a logic check against knowledge extracted from the corpus of previously published papers, conflicts could be determined already before the new experimental results are published. In analogy to the grammar- and spell-checkers, automated reasoning tools could functions as a sort of logic-checker. This way text mining could even contribute to the improvement of the semantic quality of its underlying literature corpus.

In order to include automated reasoning into the text mining workflow, however, further effort is necessary: The integration of the different knowledge representation systems was already discussed in section 11.2. Additionally, pragmatic analysis modules would probably gain more importance in order to distinguish between what is said and what is actually meant. For example, one of these analyses would be metonymy resolution. “Metonymy is a figure of speech, in which one expression is used to refer to the standard referent of a related one” (Nissim and Markert, 2003). Resolving metonymys lets one interpret sentences liek ’England won the World Cup’ in the sense that the English national football team won the World Cup instead of the country itself (Nissim and Markert, 2003). Likewise, the reconstruction of implicit information that van Dijk called MISSING LINKS (see section 11.3) would be required to draw a comprehensive picture of the extracted knowledge.

11.5 Learning to read

The focus on a controllable rule-based system in IntegreSSA might suggest that in this thesis the use of rules over machine learning methods should be promoted. The focus on rules is meant to allow improved comprehensibility and to gather more linguistically meaningful features. A rejection of machine learning, however, is not intended. Instead the extensive rule system should be seen as a comprehensive foundation for the integration of learning modules. As mentioned before, deep multi-task learning could be a way of optimizing the different syntactic, semantic and logical constraints simultaneously. Furthermore, the pattern learning modules play a central role in the prototypical sketch of a supersemantic analysis network presented in section 2.5.

These pattern learners could be the foundation for a continuously learning text mining system. In a dynamic field like science, new terms emerge constantly. Furthermore, new types of biomedical events might emerge as science proceeds and need to be identified. Different pattern learners could

accomplish these tasks. For new terms, two ways of identifying them were already presented in the course of this work. Using DefineTHAT definitions of terms could be identified. This approach could be extended in a way to automatically classify the identified terms within an ontology depending on their definition. A second way of learning term meanings is bootstrapping (explained in section 5.1). Here, terms are interpreted depending on the roles they take within known patterns. The DefineTHAT definition extraction could thus be seen as a special case of bootstrapping where only PAS patterns with a predicate of 'to be' were considered.

In addition to learning lexical meanings, more complex patterns could be learnt for keeping the event extraction of the system up to date. If the network representation of semantics is used (as proposed in section 2.5), events can be represented in the form of underspecified graph patterns as shown in section 11.2. These graph patterns in turn could be learnt at least semi-automatically by finding frequently occurring similar graphs and generalizing from them. This approach is for example used in the pattern learning phase of bootstrapping and has also been successfully applied for event extraction. For example, Liu et al. (Liu et al., 2012) identified subgraphs in dependency parses that describe biomedical events and learnt rules to generalize from these. The learnt patterns could then be manually checked and categorized as the respective event or additional learners could be applied to perform this automatically.

A typical problem of bootstrapping is the effect that errors build up over time leading to a decrease in the quality of the lexical resource as well as the pattern store. Since the two elements are used to extend each other, errors in one influence the other and this way propagate through the whole system. In the fully automatic case of a continuously learning text mining system this problem needs to be tackled. In a supersemantic system the integration of logic might prove very valuable in this connection. The logic module could constantly check the semantic graph for coherence and consistency and thereby hopefully ensure a constant quality. Such a module could work as a kind of introspection for the system to question its own choices in the past in front of the background of its accumulated knowledge.

Learning to read and to understand text is a problem that has been solved before. Humans are able to nearly flawlessly communicate with each other through language and the human brain is the machine accomplishing it. When trying to build a man-made machine being able to fullfill the same task it seems natural to orientate oneself towards already working systems. Consequently, developmental linguists try to understand how language works by investigating how it develops in children. On a different scale it might be interesting to see how the ability to use language developed within an evolutionary context. For this, studies about the ability of primates to understand language and about the development of communication systems of different species could prove valuable.

When designing a learning approach for a text mining system, the results of developmental linguists could serve as theoretical foundation. The before mentioned bootstrapping is an example where such an analogy was alread implemented. As pointed out in section 5.1, bootstrapping is inspired by the language acquisition process of children. Likewise, further evidence from developmental linguistics could be used to develop additional learning methods or refine existing ones.

For example, Waxman et al. report that in infants acquiring a language "we see a robust ability to map novel nouns to object categories, but when it comes to mapping novel verbs to event categories, a different picture emerges. Infants have considerably more difficulty. Their ability to learn the meaning of a novel verb varies as a function of the particular language they are acquiring, and within a given

language, it varies as a function of the particular linguistic contexts in which the verb appears (e.g., whether the surrounding noun phrases are mentioned explicitly or dropped)" (Waxman et al., 2013). This observation could be used to improve bootstrapping in a way that learned lexical categories are treated as more reliable than certain patterns. Furthermore, this finding suggests that this certainty estimation of bootstrapping algorithms should be adapted to the language.

Other findings focussed on the vocabulary of children during the phase of language acqusition. Tomasello noted that "their syntax was built around various particular items and expressions. [...] 92% of these children's earliest multi-word utterances emanated from one of their 25 lexically based patterns" (Tomasello, 2000). Such results support the N-gram approach to vocabulary acquisition and suggest that it might prove valuable to put such an approach first before adding additional learning modules. Further research into how these lexically based patterns are made up could provide insights on how to extract more natural N-grams.

11.6 Learning to talk

As a final step, the knowledge that is gained by a text mining analysis has to be transfered back to the human user of the text mining system. The question of how to present and represent the knowledge is a question of human-machine interaction and information science. Typically, databases or visualizations are used in this connection. A more natural way, however, might be the use of a system specifically designed for human communication - language itself. There exist a handful of fields dealing with the production of language: question answering specifically targets producing answers for freely formulated questions, automatic summarization produces summaries from a text or a corpus, and natural language generation is the more general field of any kind of language production.

The major challenge in question answering is not a linguistic one. Since answers are typically short the production of natural language answers is relatively trivial. Instead the main focus of the field is on the reasoning process that analyzes the question and relates it to the knowledge base of the system (Lopez et al., 2011). Here, it is often necessary for the program to make inferences. Thus, question answering - at least in non-trivial cases - is largely a matter of logical or statistical programming.

Automatic summarization techniques can be broadly divided into approaches that create summaries by extracting the most relevant sentences and those that create summaries by reformulating the text or corpus that is summarized. Because of its decreased complexity the former of the two is more commonly used. Here, only the relevance of a sentence has to be determined. In the latter case, however, additionally ways of generalizing and abbreviating formulations are required. This poses a major problem to summarization problems because a semantic understanding of the text is required (Nenkova and McKeown, 2011). Reformulations can be realized by using synonyms or semantically equivalent but syntactically different formulations (e.g. using an active sentence instead of a passive one). Generalizations can be realized by using a hierarchical ontology and building an analogous semantic hierarchy of verbs. Such hierarchies could then be used by replacing multiple underlying concepts by plural forms of upper levels. Take for example the following three sentences:

> Tim bought a tennis racket.
> Tim borrowed a soccer ball.

Tim purchased many golf balls.

Using an ontology that categorizes 'tennis racket', 'soccer ball' and 'golf ball' as 'sports equipment' and a verb hierarchy that identifies 'to buy', 'to borrow' and 'to purchase' as types of 'to acquire', a summarization system could summarize the three sentences as 'Tim acquired sports equipment.' Combining this approach with pruning sentence with little relevance can produce summaries with original formulations.

While inference and generalization are often complicated, the mere process of producing language is comparatively easy. In contrast to natural language understanding the problem of ambiguity vanishes. Since all entities are uniquely identifiable in the internal representation of the language generation system, there is no lexical ambiguity. Furthermore, since the structure of the sentence can be chosen, one can more easily avoid syntactic ambiguity. Consequently, most of the focus in natural language generation is on planning the structure of the conversation and topics that increase the reading ease or make the text seem more natural (Dethlefs and Cuayáhuitl, 2011). Among these are approaches to decide when to use pronouns and when to combine sentences. Both of these problems are significantly easier than there counterparts of interpreting pronouns and complicated sentence structures. Hence, the challenge of teaching a machine to talk is largely a challenge of teaching a machine to think.

12

Conclusion & Outlook

In the course of this thesis, Supersemantics were introduced as a concept of state-of-the-art text mining. Supersemantics describes a new way of doing text mining in a more integrated and less dogmatic way. In contrast to traditional approaches that separate syntactic and semantic analyses, it rejects the artificial boundaries between syntax and semantics as well as between the different fields of linguistics. Instead a practical approach that uses the best tool at the best time and thereby tries to overcome the isolation of single linguistic problems is followed. The prototypical overview of a supersemantic system given in chapter 2 can function as a blueprint for future frameworks following this paradigm. It includes linguistic analyses on the level of words, sentences, sections, texts and whole corpora and proposes methods on how to combine these with each other and external knowledge resources.

Each of these levels was presented in detail within this work. A comprehensive overview of existing methods and problems was given. In addition to this, new approaches were introduced, evaluated and discussed. The algorithms and tools implemented in this thesis systematically bridged linguistic levels as illustrated by Figure 12.1. The realized methods showed mostly very good results and could thus serve as functional modules to improve existing or future comprehensive text mining systems.

There are three ways how text mining can contribute to biological research (as also depicted in Figure 1.4): it can provide resources, help to manage and represent knowledge and produce new hypotheses. In this work, contributions of all three of these kinds were made. The first of these was successfully implemented by the generation of the Negatome 2.0 database and the ongoing development of PhenoDis, the rare diseases database. The Negatome database can be used to improve bioinformatical techniques like protein-protein interaction prediction and to detect contradictory results. The rare disease database structures information about rare diseases. This can be used as an easier to search knowledge base. Furthermore, it can be used in decision support tools and other prediction methods. The second way of text mining to contribute to biological research, the management and comprehensibility facilitation of knowledge, was tackled by the word space visualization and the text-mining-based functional

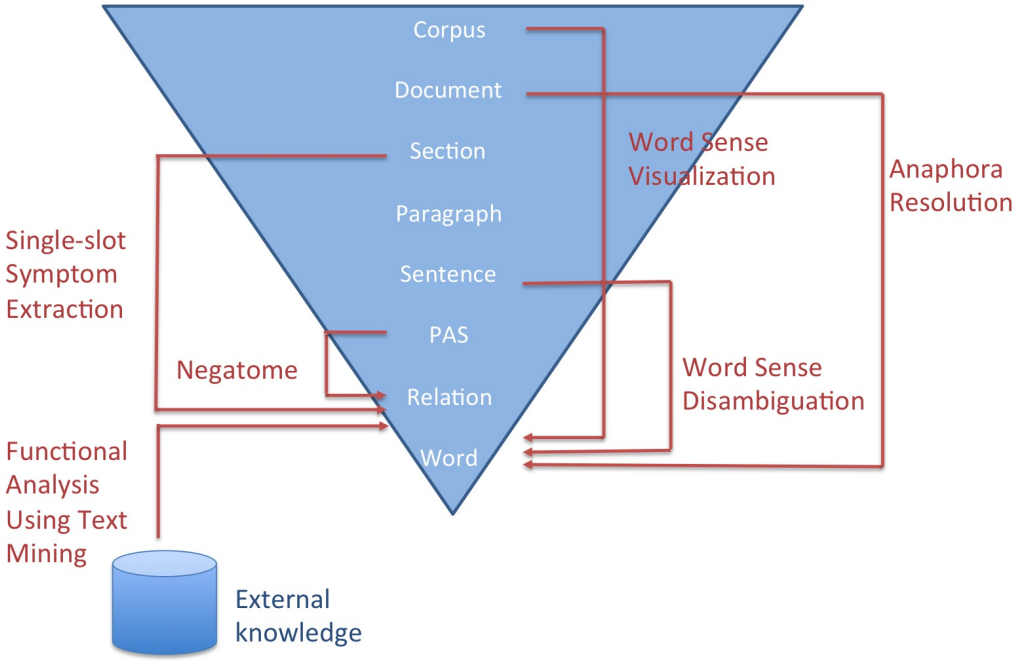


Figure 12.1: Overview of the different linguistic levels that were bridged with the different methods implemented in the course of this thesis.

analysis tool. Both tools aim at the organization of large amounts of data (derived from text mining and high throughput experiments respectively) and both tools support a researcher in making better interpretations. Finally, the third kind of contribution was the generation of hypotheses. Such a contribution was made in the form of the decision support tool in the rare diseases project. Here, the likelihood of different diseases was hypothesized based on the detected symptom profile of a patient. Furthermore, the classical ABC model can be rather easily employed on the basis of the event extraction systems Shallow SRL and IntegreSSA.

The analyses of Excerbt and Shallow SRL revealed fundamental necessities and possible pitfalls in the realization of such systems. Furthermore, the implemented prototypes showed promising first results with IntegreSSA strongly outperforming Excerbt and catching up to the performance to the longer developed leading text mining tools. Based on this, IntegreSSA might be regarded as a proof-of-concept for supersemantic analysis and as a hint on how supersemantic frameworks can be designed.

In future development, the logical next steps of this work are turning the concept as well as the developed modules and prototypes into a full-fledged supersemantic analysis framework. In this connection, the existing modules have to be improved and extended, e.g. IntegreSSA would need to provide syntactical information for the anaphora resolution system, and the ones mentioned but not yet realized, like time and location contextualization modules, would need to be added. Here, future developments concerning the width of the spectrum of provided analyses would be expected. Furthermore, often neglected analyses on superordinate levels will move stronger into focus. On text level, for example, pragmatic or discourse analysis approaches will likely be added to a supersemantic framework. Thus, text structures could be analyzed in the form of rhetoric tree structures or other text linguistic representations.

In line with the current trends of our time, it seems likely that deep learning strategies will find their way into linguistic analyses. As discussed, the multi-task deep learning approach already fits in with the multiple objectives a supersemantic system needs to satisfy simultaneously. In addition to this learning approach, further learning approaches that are more closely linked to linguistics will probably increase in importance. Using bootstrapping in combination with a logical self-correction mechanism, for example, could improve the consistency of the knowledge base, the range of ontological knowledge as well as the sentence analysis simultaneously.

Finally, with ever improving text mining systems, the range of applications will increase accordingly. Nowadays, text mining applications often require tedious preparations for the corresponding domain and for the specific question that should be answered. This includes the implementation of additional text mining modules or visualizations, the creation of vocabularies or the annotation of corpora for machine learning approaches or evaluations. With more comprehensive supersemantic systems this additional effort will decrease and the use of text mining systems will widen. With possible future applications like the support of literature research, alerts when interesting papers are published, intelligent hypothesis generation systems or open domain question answering systems, text mining applications will become a integral component of every researchers daily routine.

A

Word Sense Disambiguation Evaluation

The different training methods, preprocessing steps, and two different classifiers were evaluated for their applicability for biomedical word sense disambiguation. The results can be seen in Table A.1

Classifiers build with the CRM114 framework can be trained using the following methods:

- **train everything (TET):** All training samples are treated equally.
- **train only errors (TOE):** Only training samples, that would be classified wrongly with the classifier in its current form, are used to improve the classifier.
- **single sided thick threshold training (SSTTT):** Like TOE, but also samples that are correctly classified but under a certain confidence threshold are used (in Table A.1 SSX.X stands for SSTTT with a confidence threshold of X.X).
- **train until no errors (TUNE):** A meta-training method that is used with one of the other training methods (in Table A.1 a T is added at the end of the respective training method if it was used in combination with TUNE). Lets the training go on until all training samples are correctly classified.

If the name of a data set starts with an 'r' this means that one of the classes was reduced to contain only half of the samples. This was done to create an imbalanced data set. The Cl. column stands for the amount and type of classes used. The type is encoded in the letters that might follow the amount: G stands for gene, D for disease and O for others. Likewise, the preprocessing steps are encoded with letters: O stands for omittance of the ambigue term, R stands for replacement with a dummy term, R* means that only in the training set this replacement was performed, C stands for case folding and S stands for stemming.

Table A.1: Evaluation results of WSD system with different configurations on different data sets.

Train	Test	Cl.	Input	Classif.	Mode	Prepr.	Acc.	Prec.	Rec.	F
Manual I	T1	4	Abstract	OSB	TET	No	0.62	-	-	-
Manual I	T2	4	Abstract	OSB	TET	No	0.36	-	-	-
Manual II	T2	4	Abstract	OSB	TET	No	0.38	-	-	-
Manual II	T1	4	Abstract	OSB	TET	No	0.81	-	-	-
Manual II	T1	4	Sentence	OSB	TET	No	0.86	-	-	-
Manual II	T1	4	Sentence	OSB	TET	O	0.75	-	-	-
Manual II	T1	4	Sentence	OSB	TET	R	0.77	-	-	-
Manual II	T1	4	Sentence	OSB	TET	RC	0.74	-	-	-
Manual II	T1	4	Sentence	OSB	TET	RCS	0.75	-	-	-
Manual II	T2	4	Abstract	OSB	TET	No	0.46	-	-	-
Manual II	T2	4	Sentence	OSB	TET	No	0.48	-	-	-
Manual II	T2	4	Sentence	OSB	TET	O	0.5	-	-	-
Manual II	T2	4	Sentence	OSB	TET	R	0.5	-	-	-
Manual II	T2	4	Sentence	OSB	TET	RC	0.53	-	-	-
Manual II	T2	4	Sentence	OSB	TET	RCS	0.57	-	-	-
Manual II	T1	2 GD	Abstract	OSB	TET	No	0.91	-	-	-
Manual II	T1	2 GD	Sentence	OSB	TET	No	0.96	-	-	-
Manual II	T1	2 GD	Sentence	OSB	TET	O	0.89	-	-	-
Manual II	T1	2 GD	Sentence	OSB	TET	R	0.93	-	-	-
Manual II	T1	2 GD	Sentence	OSB	TET	RC	0.92	-	-	-
Manual II	T1	2 GD	Sentence	OSB	TET	RCS	0.92	-	-	-
Manual II	T2	2 GD	Abstract	OSB	TET	No	0.71	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TET	No	0.79	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TET	O	0.77	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TET	R	0.81	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TET	RC	0.85	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TET	RCS	0.88	-	-	-
Manual II	T2	2 GD	Abstract	OSB	TET	RCS	0.92	-	-	-
Manual II	T2	2 GD	Abstract	OSB	TOE	RCS	0.67	-	-	-
Manual II	T2	2 GD	Abstract	OSB	TOET	RCS	0.66	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS0.2	RCS	0.69	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS0.2T	RCS	0.7	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS0.5	RCS	0.7	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS0.5T	RCS	0.71	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS0.8	RCS	0.72	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS0.8T	RCS	0.74	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS1.0	RCS	0.74	-	-	-
Manual II	T2	2 GD	Abstract	OSB	SS1.0T	RCS	0.77	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TET	RCS	0.87	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TOE	RCS	0.8	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TOET	RCS	0.83	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS0.2	RCS	0.81	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS0.2T	RCS	0.84	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS0.5	RCS	0.82	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS0.5T	RCS	0.81	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS0.8	RCS	0.8	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS0.8T	RCS	0.8	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS1.0	RCS	0.82	-	-	-
Manual II	T2	2 GD	Sentence	OSB	SS1.0T	RCS	0.82	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TET	RCS	0.87	-	-	-
Manual II	T2	2 GD	Sentence	OSB	TOE	RCS	0.8	-	-	-
rMan II	T2	2 GD	Sentence	OSB	TET	RCS	0.75	-	-	-
rMan II	T2	2 GD	Sentence	OSB	TOE	RCS	0.73	-	-	-
rMan II	rT2	2 GD	Sentence	OSB	TET	RCS	0.76	-	-	-
rMan II	rT2	2 GD	Sentence	OSB	TOE	RCS	0.72	-	-	-
Manual II	Bio	2 GO	Sentence	OSB	TET	RCS	-	0.61	0.82	0.7
Manual II	Bio	2 GO	Sentence	OSB	TET	R*CS	-	0.62	0.83	0.71
Bio	Bio	2 GO	Sentence	OSB	TET	RCS	-	0.78	0.96	0.84

Bio	Bio	2 GO	Sentence	OSB	TET	R*CS	-	0.78	0.92	0.84
Bio	Bio	2 GO	Sentence	OSB	TET	CS	-	0.52	1.0	0.68
Bio	Bio	2 GO	Sentence	MV	TET	R*CS	-	0.8	0.91	0.85

B

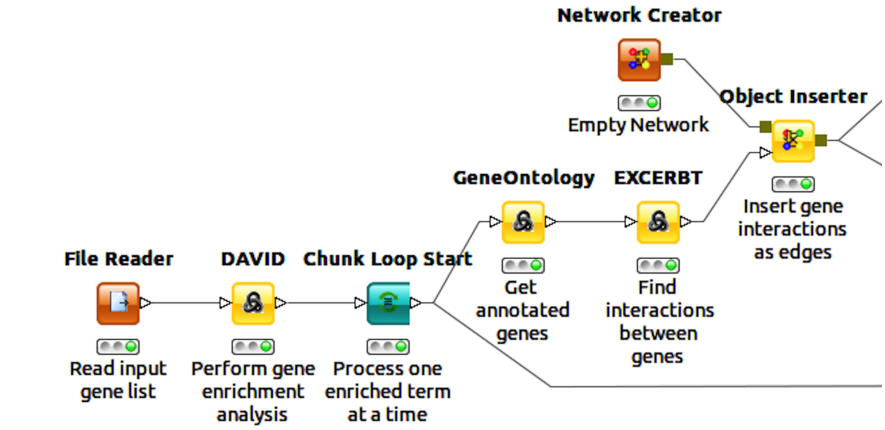
Modular implementation of functional analysis tool

Software that was developed in the course of scientific works is rather rarely reused and in even fewer cases parts of the implementation are integrated into other programs. In order to counteract this trend some developers publish their code under open source licenses. However, rarely modular approaches with standardized interfaces are used in order to simplify the integration of parts of the software. One way of standardization is to implement ones programs as modules in general workbenches for data analysis like the Konstanz Information Miner (KNIME) (Berthold et al., 2006). Such an approach was taken for the functional analysis tool described in chapter 8. The KNIME integration is described in this appendix.

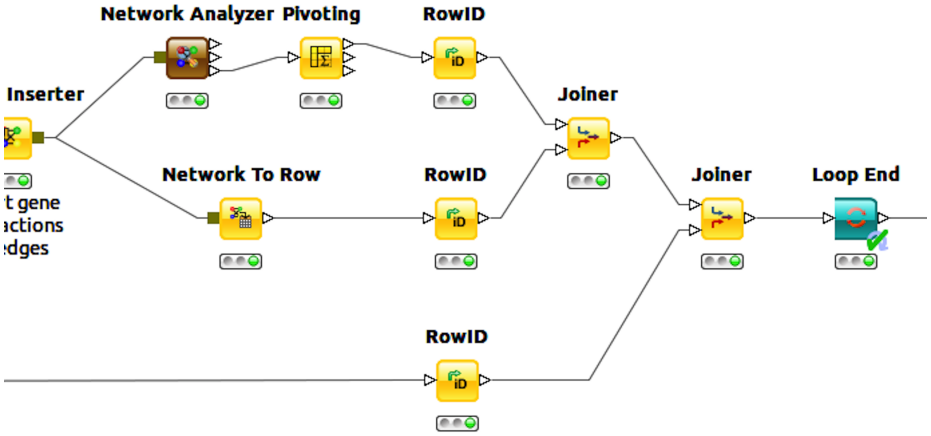
KNIME was developed at the University of Konstanz. It offers an environment in which a user can easy build analysis pipelines (so-called workflows) using an intuitive visual representation. The steps in the pipeline are implemented as so-called nodes. The different nodes usually communicate between each other via tables. This simple format allows the quick integration of various different analysis modules. Furthermore, the basic node repository of KNIME already offers a variety of nodes that one can immediately integrate into ones own pipelines. This includes nodes for I/O, statistical calculations, data mining algorithms, and visualizations (Berthold et al., 2006, 2008; KNIME development team, 2013).

In the course of this work the code for the communication with DAVID, the Gene Ontology database, Excerbt and String were implemented as KNIME nodes. Since all of these are established resources, this should provided a high reusability of the components of the program. Furthermore, two workflows (one including Excerbt, one including String) were designed that are able to perform the whole analysis from reading the input to visualizing the graphs. An overview of one of these workflows (the one containing Excerbt) is given in Figure B.1.

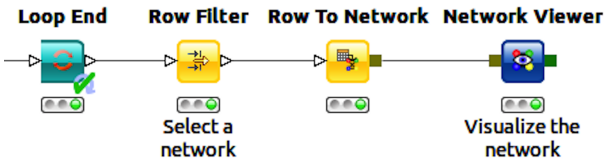
As can be seen, the gene list is first read using a File Reader node. Then DAVID, the GO and Excerbt are called. Using the Network Creator node this information is combined into a network (a). The network is



(a) Generation of interaction network



(b) Calculation of network measures



(c) Visualization

Figure B.1: KNIME workflow for text mining based functional analysis. Picture taken from (Jeske, 2013).

then analyzed using the Network Analyzer node (b). Finally, the network is visualized with the Network Viewer node (c).

C

Additional results of GO analysis

Section 8.3 described a graph theoretic analysis of the Gene Ontology. This appendix supplements this section by providing additional results. Figure C.1 shows how the investigated graph measures change if a different confidence score is used. The plots show the average density (a), average clustering coefficient (b), average portion of the largest component in the graph (c) and average diamater (d) with respect to the depth in the hierarchy of the GO.

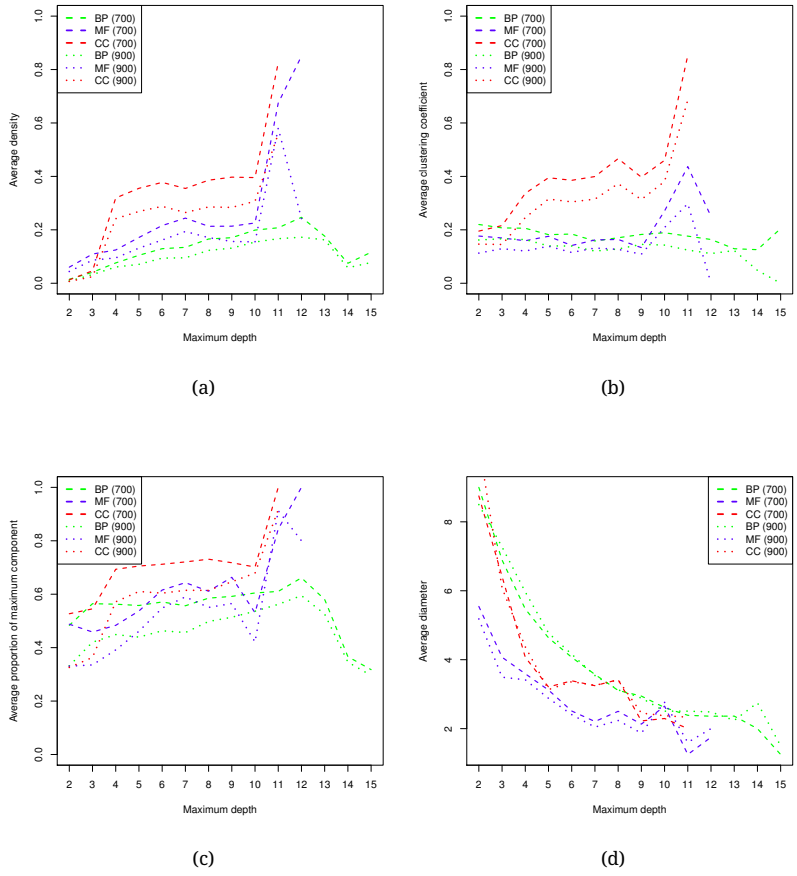


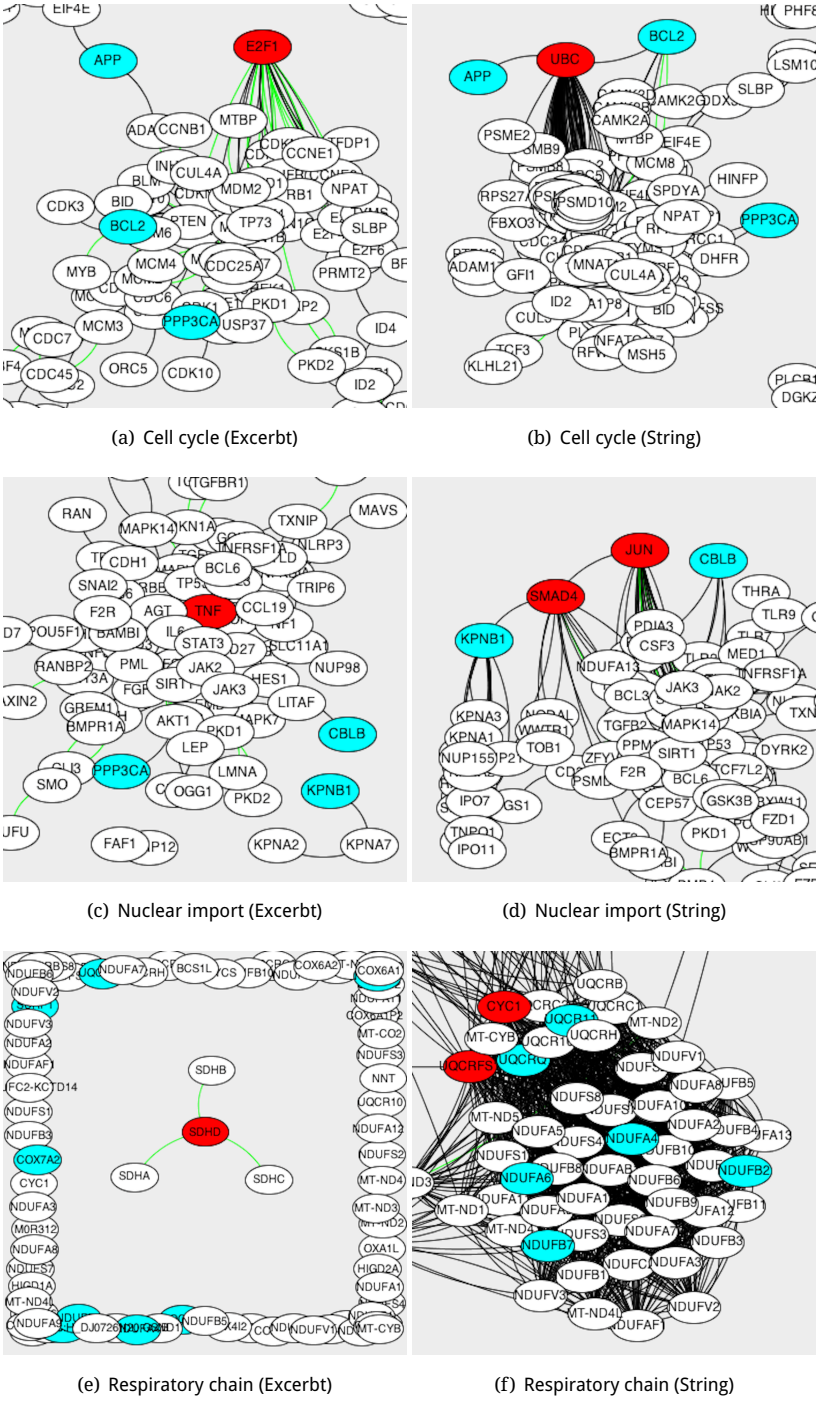
Figure C.1: Comparison of density, average clustering coefficient, proportion of the largest component of the graph and diameter for different confidence scores of String.

D

Additional results of functional analysis of epilepsy study

In section 8.4 the application of a text mining based functional analysis to a gene set enrichment study by Greiner et al. (Greiner et al., 2013) was described. Figure D.1 shows the graphs for some of the most relevant GO terms with respect to generalized seizure epilepsy patients from this study. Both the graphs based on interactions found by Excerbt and String are shown for the GO terms 'cell cycle', 'nuclear import' and 'respiratory chain'.

As can be seen, Excerbt and String detect different central genes in all GO categories. Interestingly, for the cell cycle String finds two connections from genes from the enriched gene list to the central gene UBC. Since UBC regulates the cell cycle, this might be an indicator for a very direct connection in patients with generalized seizures. Excerbt, on the other side, neglects UBC due to the missing resolution of synonyms. Likewise, for the nuclear import String finds direct connections between genes of the input list and a central gene, while Excerbt does not. This together with the completely diverging results for the respiratory chain shows that the results can differ rather largely depending the used text mining resource.



(a) Cell cycle (Excerbt)

(b) Cell cycle (String)

(c) Nuclear import (Excerbt)

(d) Nuclear import (String)

(e) Respiratory chain (Excerbt)

(f) Respiratory chain (String)

Figure D.1: Text mining based graphs of GO term respiratory chain for generalized seizure epilepsy patients.

E

Measuring semantic entropy of texts

Readibility Measures

The evaluation and interpretation of texts is often a very subjective process. As a consequence of this the variation in how texts are recepted is rather large. In many situations, e.g. when judging the quality of essays in school or university, however, objective comparability would be desired. One approach that tries to objectify texts is the application of different text measures. By providing universally valid and automatically calculatable measures the variation in interpretation is reduced (at least with respect to the aspects covered by the measures).

The main focus in science when developing such measures are those that try to determine the readibility of texts. In the 1940s, different such measures were introduced. Two of the most well-known ones were the Dale-Chall formula(Dale and Chall, 1948) and the Flesch reading ease formula (Flesch, 1948). The Dale-Chall formula (DCF) is defined as follows:

$$DCF = 0.1579 * 100 * \frac{d}{w} + 0.0496 \frac{w}{s} \tag{E.1}$$

Here, w is the amount of words in the text, s is the amount of sentences and d is the amount of difficult words. This last value can be estimated by using a stop-word list and considering all words that are not on that list as difficult. Consequently, the exact value of the Dale-Chall formula always depends on the used stop-word list. Originally, a list containing 763 words that 80% of fourth-grade students were familiar with was used. The DCF is accompanied by a table (see Table E.1) that can be used to interpret the values.

While the Dale-Chall formula focusses on the frequency of the used words. The Flesch reading ease formula focusses on the complexity of the used words and sentences. For this purpose, the number of

Table E.1: Interpretation of values calculated with the Dale-Chall formula. Based on (Dale and Chall, 1948).

Value	Interpretation
$\leqslant$4.9	easily understandable for students from 4th grade or lower
5.0 - 5.9	easily understandable for students from 5th or 6th grade
6.0 - 6.9	easily understandable for students from 7th or 8th grade
7.0 - 7.9	easily understandable for students from 9th or 10th grade
8.0 - 8.9	easily understandable for students from 11th or 12th grade
9.0 - 9.9	easily understandable for students from 13th to 15th grade (college)
$\geqslant$ 10.0	easily understandable for college graduates

words per sentence *ASL* and the number of syllables per word *ASW* is considered. The reading ease formula is defined as follows:

$$FRE = 206.835 - (1.015 * ASL) - (84.6 * ASW) \tag{E.2}$$

The interpretation of the values calculated by this formula is given in Table E.2.

Table E.2: Interpretation of values calculated with the Flesch reading ease formula. Based on (Flesch, 1948).

Value	Interpretation
0-30	Very difficult, for academics
30-50	Difficult
50-60	Moderately difficult
60-70	Moderate, for 13-15 year old pupils
70-80	Moderately easy
80-90	Easy
90-100	Very easy, for 11 year old pupils

BlabberTracker

In the course of this work an additional text measure was developed. While the existing formulas commonly focussed on rather formal syntactical features, this measure tries to tackle the semantic information content of a text. The resulting measure is integrated into a tool called BlabberTracker that tries to distinguish precise texts with a high semantic information content from those that talk a lot without saying much.

The information content of messages is a well established measurand in the field of information science. The most common method of acquiring this value is by calculating the Shannon entropy of the message. This entropy $E(X)$ is defined as follows:

$$E(X) = -\sum_{x_i \in X} P(x_i) log(P(x_i)) \tag{E.3}$$

In information science, the entropy is frequently calculated to determine e.g. the possible compression rate. It then denotes how long the message needed to be in order to transport all information if the best possible representation would have been used. In this connection the entropy of message X is calculated on its single characters x_i. But the entropy could also be calculated based on words or other elements of a text.

The train of thought followed in the development of the BlabberTracker measure is to use the entropy on word level to only focus on words that transport semantic meaning. In linguistics, words can be categorized in two classes: open class words and closed class words. Open class words are those that carry semantic meaning. Adjectives, verbs, nouns and adverbs fall into this category. Closed class words on the other hand are those that determine the syntactic structure of a sentence but do not contribute any meaning on their own. Words in this category are among others determiners, prepositions and conjunctions. Thus, in order to focus on the semantically meaningful utterances the Shannon entropy is calculated only after filtering out all closed class words. The remaining words are used in a bag of words representation to calculate the entropy on them.

One problem that occurs when dealing with text measures is the danger of creating a measure that is overly dependent on the length of the text. While a complete elimination of such effects is often impossible at least a minimization of them is desired. In order to achieve this, a normalization procedure was applied that tries to counteract text length dependencies. The positive effect of such normalization approaches has been shown in many cases (see e.g. (Singhal et al., 1996)). Since the semantic entropy is calculated on open class words, the normalization used here is likewise defined on open class words:

$$E(X) = -\frac{\sum_{x_i \in X} P(x_i) log(P(x_i))}{1 + \alpha * log(t)} \tag{E.4}$$

Here, t is the total amount of open class words in the given text. The logarithmic value of t is used for the normalization. The one is added to receive positive values. Furthermore, the value is scaled by a factor α to soften the effect. In the results here α=0.4 was chosen.

For better comparability the Dale-Chall formula and the reading ease formula were additionally implemented. The different results obtained by the three measures can be seen in the following section. In order to provide a most comprehensive tool the BlabberTracker also includes the two readability measures.

Results

As mentioned before, the very objective of text measures is to provide an objective assessment of texts that is missing so far. Thus, there does not exist a gold standard for a measure like the one presented here. In order to still provide a certain level of comparability, the obtained results are compared

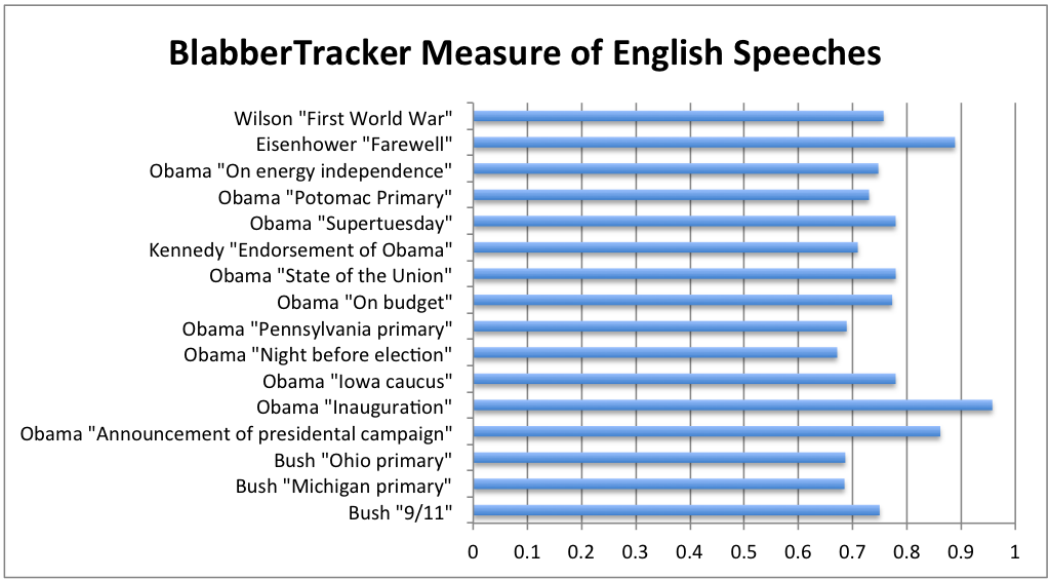


Figure E.1: BlabberTracker measure values of different speeches of American politicians.

with the Dale-Chall formula and the reading ease formula. Furthermore, it is shown what values are obtained for different speeches of politicians. One would expect that more important and better perceived speech should have higher levels of semantic information content.

For evaluation of the measure a collection containing important speeches of the most recent presidents George W. Bush (on the terror attacks of 9/11) and Barack Obama (press conference about the death of Bin Laden, a state of the union, a budget speech) as well as several speeches during their presidental campaigns were chosen. Furthermore, two historically important speeches of the former presidents Eisenhower and Wilson were included.

The results of BlabberTracker measures is shown in Figure E.1. As can be seen, Obama's inauguration speech, his announcement of to run for president and Eisenhower's farewell address reach the highest values. In his final speech Eisenhower warned the public about the military-industrial complex. He tried to convey a clear and urgent message. Thus, high values of semantic content would be expected here. Furthermore, Obamas most important speeches stick out compared to all the other considered speeches he held during his campaign. This also conforms with the expectations of what a semantic content measure should capture. The same picture is seen within the speeches of George W. Bush. His address to the 9/11 attacks scores higher than his campaign speeches. However, this last effect is rather small compared to the differences between different speakers. The lowest scoring speech is Obama's address on the last night before the election. At this point in the campaign it is unlikely that further information should be communicated but rather that his supporters should be motivated and stimulated to vote. Thus, here rather empty phrases and paroles would be expected. The low BlabberTracker score confirms this consideration.

Additionally, an evaluation of German speeches was conducted in order to see whether there exist language dependent effects. The collection of German speeches also includes important historical speeches. Here, the propaganda speeches of Hitler and Goebbels as well as an important oppositional

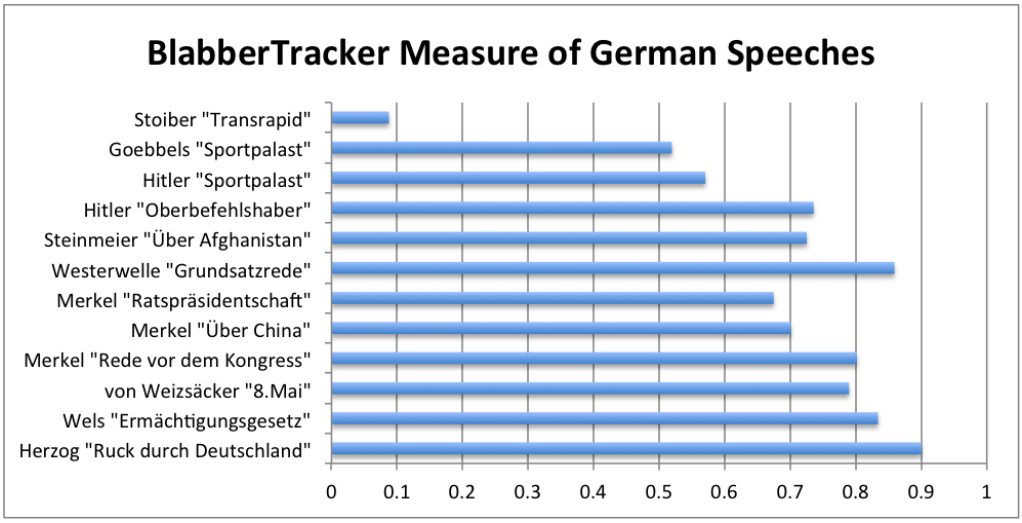


Figure E.2: BlabberTracker measure values of different German speeches.

speech by Otto Wels were included. Furthermore important speeches by former presidents Herzog and von Weizsäcker were chosen as well as speeches of contemporary politicians like Merkel, Westerwelle and Steinmeier. Finally, Edmund Stoiber's infamous gibberish about the planned installation of the Transrapid in Munich was included as a negative control.

The results obtained on German texts can be seen in Figure E.2. The picture of the German texts seems even more clear than the English one. The confusing speech of Stoiber scores by far lowest. Furthermore, propaganda speeches have low values, while important speeches like Herzog's "Ruck durch Deutschland" are at the top of the list. A surprisingly high value is also scored by Westerwelle's keynote address.

As can be observed, for both languages the supposedly qualitatively higher speeches on average obtained higher scores. While further evaluations seem necessary, the first results indicate that the BlabberTracker measure produces reasonable results. Within the scientific community it might be of value since it can help create better and more informative texts. A possible application of it could be to use it as a check before publishing publications, books or scripts for students.

For comparision, Figures E.3 - E.5 show the results of the Dale-Chall formula and the Flesch reading ease. The reading ease formula was only implemented for English due to a lack of an algorithm to calculate the syllable count in German texts. As can be seen, the Dale-Chall measure gives overall rather large values. The highest scores are Obama's budget and energy indepence speeches as well as Wilson's and Eisenhower's speeches. Obama's campaign speeches score lower. In German, two propaganda speeches of Hitler and Goebbels score highest. More technical speeches like the ones of Merkel on China and Steinmeier on Afghanistan score lowest. The values do not differ much and the results in English and German seem to be inconsistent (high values for propaganda in one and for meaningful ones in the other). Here, the BlabberTracker seems to work better to distinguish semantic content.

The results of the Flesch formula give a more comprehensive picture. The speeches of Wilson and Eisenhower stand out from the other speeches by having significantly lower values. The highest scores,

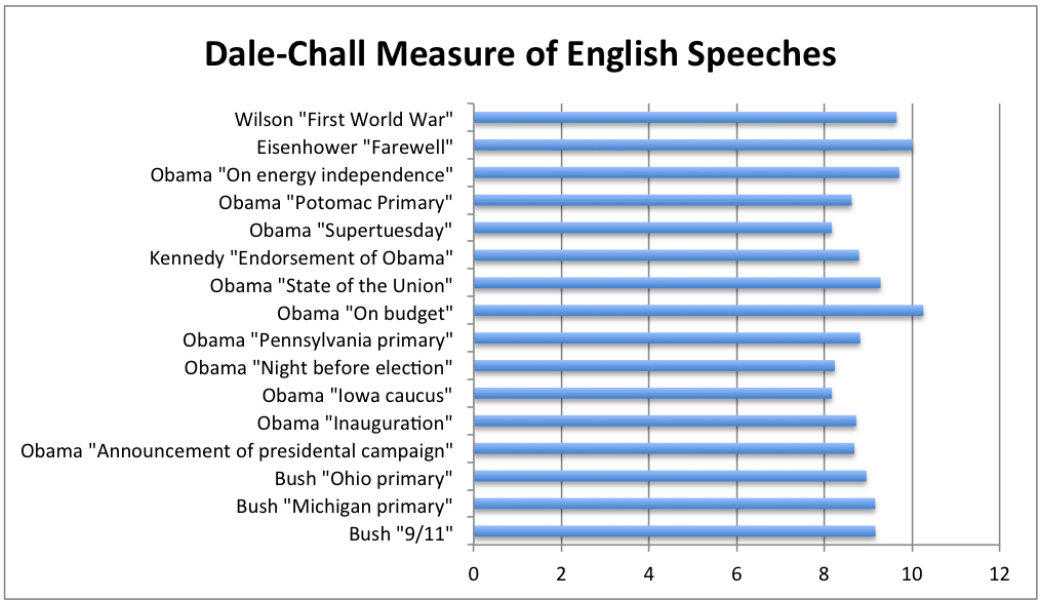


Figure E.3: Dale-Chall formula values of different speeches of American politicians.

on the other hand, are achieved by different campaign speeches. Thus, the Flesch readibility formula seems to be a good complement for the BlabberTracker measure.

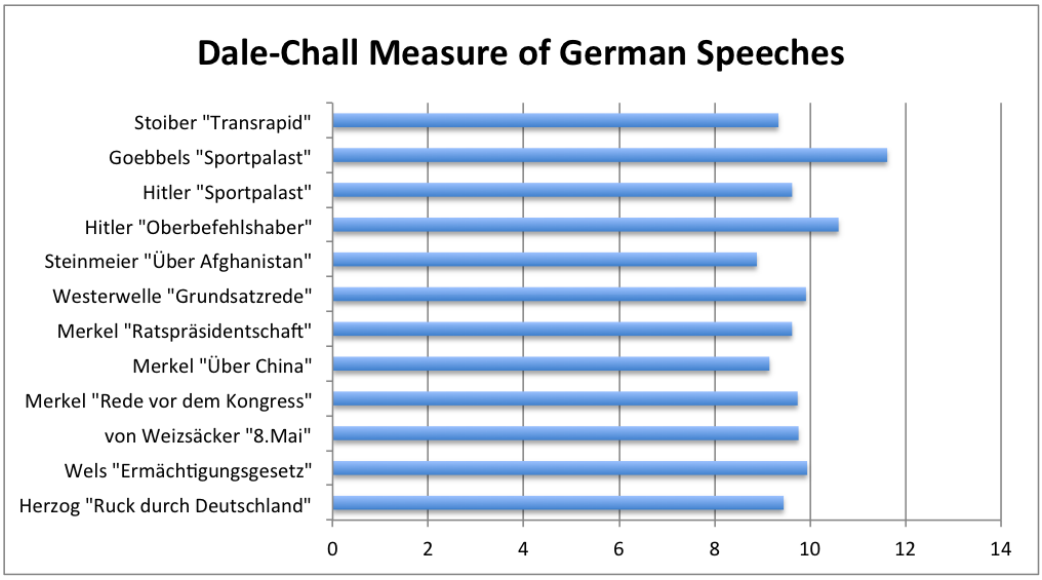


Figure E.4: Dale-Chall formula values of different speeches of German politicians.

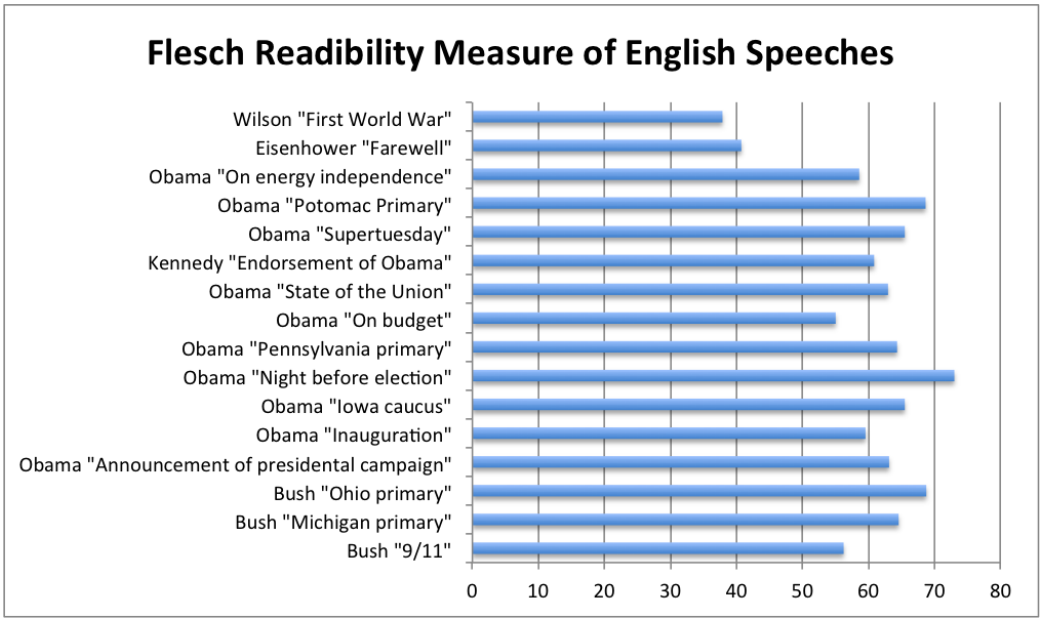


Figure E.5: Flesch reading ease values of different speeches of American politicians.

F

POS tag set used in German IntegreSSA

Table F.1: STSS POS tag set. Taken from (Sappelt, 2013), which is in turn based on (Hirschmann, 2014; Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2013). The STSS POS tag is used in the Tiger corpus and also in the POS tagger used in German IntegreSSA.

POS tag	Description	Example
ADJA	Attributive adjective	eine **latente** Hypertonie
ADJD	Adverbial or predicative adjective	er fährt **schnell**; er ist **schnell**
ADV	Adverb	sie kommt **bald**
APPR	Preposition; left part of circumposition	**nach** Berlin
APPRART	Preposition with article	**zur** Sache
APPO	Postposition	der Sache **wegen**
APZR	Right part of circumposition	von mir **aus**
ART	Definite or indefinite article	**ein** Haus; **die** Ärztin
CARD	Cardinal number	**zwei** Männer; im Jahre **1994**
FM	Foreign word	er hat das mit **“a big fish”** übersetzt
INTJ	Interjection	**ach, tja** dann halt nicht
KOUS	Subordinating conjunction	sie wartet, **weil** sie früh dran ist
KOUI	Subordinating conjuction with ’zu’-infinitive	**um** zu arbeiten
KON	Coordinating conjunction	sie **und** Emma warten **und** lesen
KOKOM	Comparative conjunction	schneller **als** er; schneller **wie** er
NN	Noun	der **Computer**; die **Patientin**
NE	Proper noun	**Hans**; **Hamburg**; **Diabetes**
PDAT	Attribute-adding demonstrative pronoun	**jene** Männer; **dieses** Spanisch
PDS	Substituting demonstrative pronoun	**denen** war **dies** nicht übelzunehmen
PIAT	Attribute-adding indefinite pronoun	**kein** Mensch; **irgendein** Glas
PIS	Substituting indefinite pronoun	**keiner**; **viele**; **man**; **niemand**
PPER	Non-reflexive personal pronoun	**ich**; **er**; **mich**; **ihm**
PPOSAT	Attribute-adding possessive pronoun	**mein** Buch; seine **Mutter**
PPOSS	Substituting possessive pronoun	**meiner**; **deines**
PRELS	Substituting relative pronoun	der Hund, **der**
PRELAT	Attribute-adding relative pronoun	der Mann, **dessen** Hund
PRF	Reflexive personal pronoun	**sich**; **einander**; **dich**; **mir**
PWS	Substituting interrogative pronoun	**wer**, **was**
PWAT	Attribute-adding interrogative pronoun	**wessen** Hut, **welche** Farbe

PWAV	Adverbial interrogative pronoun	**warum**; **wo**; **wann**; **worüber**; **wobei**
PROAV	Pronominal adverb	**dafür**; **dabei**; **deswegen**; **trotzdem**
PTKZU	'Zu' before infinitive	**zu** gehen
PTKNEG	Negation particle	**nicht**
PTKVZ	Particle part of separable verb	Er kommt **an**; er fährt **rad**
PTKANT	Answer particle	**ja**; **nein**; **danke**; **bitte**
PTKA	Particle 'am'/'zu' before adjective/adverb	**am** schnellsten; **zu** teuer
TRUNC	Detached first part of compound noun	**An-** und Abreise
VVFIN	Finite full verb	du **gehst**; wir **kommen** an
VAFIN	Finite auxilliary verb	du **bist**; wir **werden**
VMFIN	Finite modal verb	du **darfst**; sie **sollte**
VVINF	Infinite full verb	**gehen**; **abreisen**
VAINF	Infinite auxilliary verb	**sein**; **werden**
VMINF	Infinite modal verb	**wollen**; **sollen**
VVIMP	Imperative full verb	**geh**!; **reise** ab!
VAIMP	Imperative auxilliary verb	**sei**!; **werde**!
VVPP	Past participle of full verb	**gegangen**; **abgereist**
VAPP	Past participle of auxilliary verb	**gewesen**
VMPP	Past participle of modal verb	**gekonnt**; er hat gehen **können**
VVIZU	Full or particle erb in 'zu' infinitive	**wegzuhören**
XY	Non-word, special character, abbreviation	es enthält viel **D2XW3**
$,	Comma	**,**
$(	Other sentence-internal punctuation	**( )**
$.	Sentence-ending punctuation	**. ! ? ; :**

G

Additional example of chunking with German IntegreSSA

In German IntegreSSA, special levels for chunking for the German language were implemented. The results of the sentence analysis using these levels can be seen in Figure G.1.

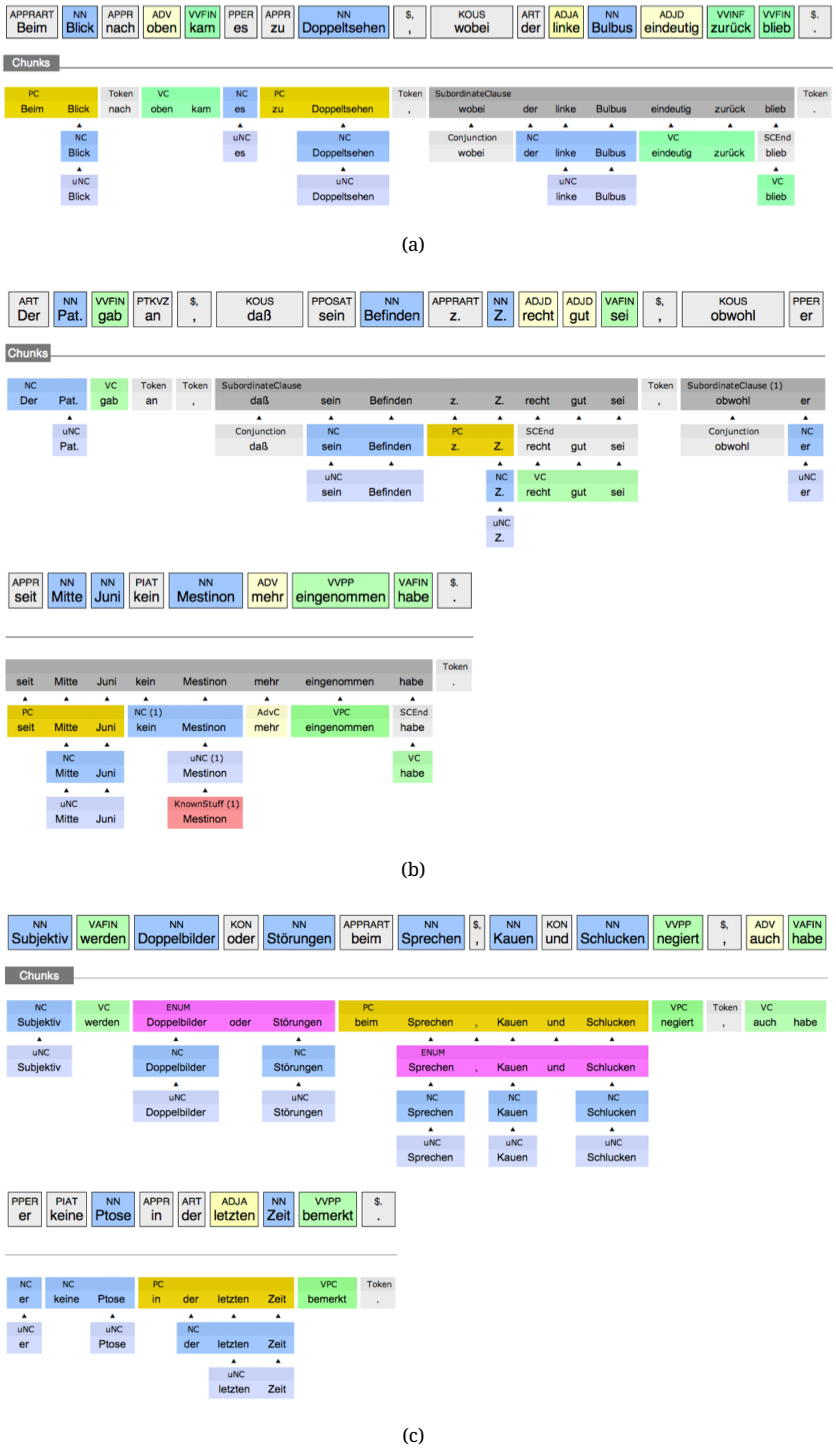


Figure G.1: Results of the German chunking analysis for three randomly chosen sentences.

List of Abbreviations

AI	Artificial intelligence
BCR	Bayesian chain rule
CFG	Context-free grammar
CIQ	Chunk in question
COPD	Chronic obstructive pulmonary disease
DAVID	Database for Annotation, Visualization and Integrated Discovery
DCF	Dale-Chall formula
DT	Determiner
ER	Event recognition
GO	Gene Ontology
GUI	Graphical user interface
GvC	Generalized seizure vs. control group
HPO	Human Phenotype Ontology
IE	Information Extraction
IntegreSSA	Integrated supersemantic analysis
IPF	Idiopathic pulmonary fibrosis
IR	Information Retrieval

JJ Adjective

KEGG Kyoto Encyclopedia of Genes and Genomes

LSA Latent semantic analysis

LSI Latent semantic indexing

ML Machine learning

MPO Mammalian Phenotype Ontology

MV Markovian classifier

NaCTeM National Centre for Text Mining

NC Noun chunk

NE/PR Named entity/pattern recognition

NER Named entity recognition

NIP Non-interacting protein pair

NLP Natural language processing

NN Noun, singular or mass

NP Noun phrase

ORDR Office of Rare Diseases Research

OSB Orthogonal sparse bigram classifier

OSBF Orthogonal sparse bigram classifier with frequency features

PAS Predicate-argument structure

PC Prepositional chunk

POS Part-of-speech

PP Prepositional phrase

PPI Protein-protein interaction

PvC Partial seizure vs. control group

RI Random indexing

RR	Relation recognition
SBPH	Sparse binary polynomial hashing
SRDD	Swedish Rare Diseases Database
SRL	Semantic role labeling
SSTTT	Single sided thick threshold training
SVD	Singular value decomposition
TEES	Turku event extraction system
TET	Train everything
TM	Text mining
TOE	Train only errors
TUNE	Train until no errors
WSD	Word sense disambiguation

List of Figures

1.1 The diagram shows the growth of the Medical Literature Analysis and Retrieval System Online (MEDLINE), the main part of PubMed, the most comprehensive collection of biomedical knowledge, in the years between 1965 and 2012 (Pubmed, 2014). 3
1.2 Hierarchy of the different levels of biological complexity in the domain of plant biology. Figure taken from (Trewavas, 2006). 6
1.3 Hypothesis-driven loop of systems biology. Figure taken from (Kitano, 2002). 7
1.4 The workflow of a prototypical text mining system. Document are fist collected using information retrieval methods and then analyzed using information extraction methods. The structured information is then used to create new hypotheses or directly stored in data bases or used to build other tools. The depiction of the hypothesis generation step is taken from (Evans and Rzhetsky, 2010). 9
1.5 An example sentence to show the high level of ambiguity in gene names and aliases. All words in this sentence are names or aliases of genes or gene products. 11
1.6 Overview of the components of an Information Retrieval system. Figure based on (Larson, 2012). 14
1.7 Prototypical representation of an information extraction pipeline. 16
1.8 The text mining landscape. Some of the most influential research groups and researchers are shown. 19
1.9 Simplified schematic description of different approaches to text mining. 20
1.10 Network of some of the most influential entities of the biomedical text mining community. 21

2.1 Different levels of linguistic utterances. The arrows indicate the interconnections between them. While classical approach usually only address the bottom-up direction, supersemantics tries to take both into account. 25
2.2 Different levels of linguistic utterances and how they need to be bridged in order to solve typical linguistic problems (on the right). Furthermore, some linguistic disciplines that focus largely exclusively on one of the levels are shown (on the left). 26
2.3 Different levels of abbreviations of the term V-SNARE. Figure based on (Morgan, 2005). . 32
2.4 A prototypical overview of a comprehensive supersemantic analysis network. 35

3.1 Workflow of the biomedical WSD system. Picture taken from (Winkler, 2011). 42

3.2 Example feature selection process using SBPH. Picture taken from manage-this.com [15]. . 44
3.3 Comparison of WSD-based NER with contastants of BioCreAtIvE task 1A. The results for all 15 participants of the BioCreAtIvE task (A-O) for the open and closed version of the task are shown. 46

4.1 Tool for annotating the non-interactions proposed by Excerbt. 53
4.2 Acceptance Rate of Negatome annotation. One can see what ratio of Excerbt proposals was accepted by the manual curator and how many non-interactions the annotator added from papers proposed by Excerbt. 54

5.1 Simplified version of an information extraction pipeline that can be used for relation extraction in single-slot tasks. 62
5.2 Average amount of entities found with NER using all vocabularies. 64
5.3 Average amount of symptoms found with NER using different vocabularies. 65
5.4 Performance of the decision support tool with and without weighting for different cutoffs relative to the total amount of diseases. 66

7.1 Visualization of Excerbt text mining results of the term 'proteasome'. 84
7.2 Overview of the architecture of the visualization tool. 85
7.3 Multi-concept mode of the visualization tool. 86
7.4 Visualization of the text mined results for IPF. 87
7.5 Visualization of the text mined results for COPD. 89

8.1 Overview of the workflow of the text mining assisted functional analysis tool. Picture taken from (Jeske, 2013). 95
8.2 Input window of the tool. Picture taken from (Jeske, 2013). 96
8.3 Overview window displaying the highest three levels of the GO. Only the enriched GO terms are shown. The color indicates the degree of enrichment with darker terms being more enriched. Picture taken from (Jeske, 2013). 97
8.4 Screenshot of the main window of the functional analysis tool. Picture taken from (Jeske, 2013). 98
8.5 Graph measures of GO categories by depth within the GO hierarchy. Pictures taken from (Jeske, 2013). 100
8.6 Text mining based graphs of GO terms relevant for partial seizure epilepsy patients. Pictures taken from (Jeske, 2013). 103
8.7 Text mining based graphs of GO term respiratory chain for generalized seizure epilepsy patients. Pictures taken from (Jeske, 2013). 105

9.1 Shallow SRL processing pipeline. 114

10.1 Schematic overview of IntegreSSA 125
10.2 Example output of the IntegreSSA sentence analysis. 127
10.3 Processing times of Senna in relation to the length of a sentence. 132
10.4 Processing times of the IntegreSSA sentence analysis in relation to the length of a sentence. 132
10.5 Example output of the German IntegreSSA chunking analysis. 138

11.1 Prototypical architecture of a multi-task learning system that learns representative factors for each subtask and thereby strengthens the generalization abilities of the model. Figure taken from (Bengio et al., 2013). 144
11.2 Overview of different approaches to knowledge reprentation. 147
11.3 Sketch of a possible data structure integrating syntactic, semantic, logical and statistical information. 148
11.4 Sketch of a possible graph pattern to represent the meaning of the term 'inhibitor'. 149

12.1 Overview of the different linguistic levels that were bridged with the different methods implemented in the course of this thesis. 160

B.1 KNIME workflow for text mining based functional analysis. Picture taken from (Jeske, 2013). 168

C.1 Comparison of density, average clustering coefficient, proportion of the largest component of the graph and diameter for different confidence scores of String. 172

D.1 Text mining based graphs of GO term respiratory chain for generalized seizure epilepsy patients. 174

E.1 BlabberTracker measure values of different speeches of American politicians. 178
E.2 BlabberTracker measure values of different German speeches. 179
E.3 Dale-Chall formula values of different speeches of American politicians. 180
E.4 Dale-Chall formula values of different speeches of German politicians. 181
E.5 Flesch reading ease values of different speeches of American politicians. 181

G.1 Results of the German chunking analysis for three randomly chosen sentences. 186

List of Tables

1.1 The roles assigned by the Senna role labeling tool. The roles correspond to the Propbank definition. The example sentences are based on or taken from the examples from the Propbank annotation guidelines Bonial et al. (2010). 18

2.1 Supersemantics: Related work 37

3.1 Word Sense Disambiguation: Related work 47

4.1 PAS Contextualization: Related work 56

5.1 NER Performance on sample of GeneReviews with different vocabularies. 65
5.2 Evaluation of decision support tool. 66
5.3 Decision Support Systems: Related work 67

6.1 Performance of anaphora resolution system on BioNLP data. 74
6.2 Text contextualization: Related work 77

7.1 Association scores for different word space approaches. 88
7.2 Corpus contextualization: Related work 91

8.1 Text mining-based functional analysis: Related work 105

9.1 Utterances that describe an event without using a verb. 110
9.2 Chunk types of the OpenNLP chunker. The type set is based on CoNLL-2000 shared task (Sang and Buchholz, 2000). The examples are mostly taken from Sang and Buchholz (Sang and Buchholz, 2000). 115
9.3 Overview over the features used by the support vector machine in Shallow SRL. 117
9.4 Related work on automatic definition detection. 120

10.1 Performance of IntegreSSA in comparison to Excerbt on BioNLP shared task GE task data. 134
10.2 Units supported by the German IntegreSSA measurement recognizer. 138
10.3 Chunk types used by German IntegreSSA. 139

10.4 Results of the official BioNLP shared task GE task (Nédellec et al., 2013) submissions for core event extraction. 141

A.1 Evaluation results of WSD system with different configurations on different data sets. . . 164

E.1 Interpretation of values calculated with the Dale-Chall formula. Based on (Dale and Chall, 1948). 176
E.2 Interpretation of values calculated with the Flesch reading ease formula. Based on (Flesch, 1948). 176

F.1 STSS POS tag set. Taken from (Sappelt, 2013), which is in turn based on (Hirschmann, 2014; Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2013). The STSS POS tag is used in the Tiger corpus and also in the POS tagger used in German IntegreSSA. 183

Glossary

Anaphora

is a referential utterance pointing to another utterance, the antecedent. Since the anaphora is only the reference for interpreting its meaning the antecedent is required. Typical anaphoras are e.g. pronouns. 15, 29, 55, 71, 122, 136, 161

ARG0

is the argument referring to the active entity that initializes the event described by a predicate-argument-structure. 17, 18, 51, 52, 73, 109, 116, 119, 133

ARG1

is the argument referring to the passive entity that is affected by the event described by a predicate-argument structure. 17, 18, 51, 52, 109, 116, 119, 123, 133

Association Patterns

are frequently occurring patterns that can be used to improve linguistic analysis on several levels. If e.g. an association pattern of a typical formulation “as mentioned by x” where x is always a person is found, then this can be used to distinguish a meaning of a word that describes a person from another one. 36, 79

Co-occurrences

is a text mining procedure in which relations between entities are inferred based on the frequency with which they occur within close proximity to each other. 15, 20, 50, 80

Constituency Parser

is an linguistic procedure that analyzes the structure of a sentence using a grammar and represents it in the form of a parse tree containing non-terminal grammar categories and terminal POS tags. 12

Dependency Parser

is an linguistic procedure that analyzes the structure of a sentence and represents it by arranging the words of the sentence in a parse tree. 12

Event

is a relation or other linguistic pattern that is specifically searched for during the information extraction stage of a text mining system. In the biomedical domain typical events include gene expression and positive or negative gene regulation. 14, 17, 20, 34, 76, 110, 111, 122, 123, 131

Excerbt

is a Senna-based text mining system that extracts binary relations between a multitude of entities. 17, 46, 51, 53, 73, 80, 85, 96, 107, 122, 144, 161

Information Extraction

is the scientific field that tries to extract knowledge in structured form (typically so-called entities and relations) from unstructured texts using computer-aided methods from artificial intelligence and computer linguistics. 9, 14, 57

Information Retrieval

is the scientific field that deals with the computer-aided search for information that satisfies a certain information need from a given collection of documents, songs, videos or any other information bearing resource. 8, 13, 70

Named Entity Recognition

is a linguistic procedure that associates words in a text with their corresponding entities in a knowledge base. Typical named entity recognizers find persons, companies, or in the biomedical domain proteins or diseases. 14, 20, 34, 45, 61, 123, 124, 140

Natural Language Processing

is the scientific field that processing language using methods from artificial intelligence and computer linguistics to automatically solve a variety of linguistic problems. 9, 71, 115, 123, 128, 144

Part-of-speech tagging

is the linguistic procedure of assigning so-called parts-of-speech to tokens. Parts-of-speech are finer grained word classes like gerund verbs or plural nouns. 12, 16, 34, 76, 80, 115, 123, 124, 128, 136, 139

Predicate-argument-structure

is a representation of semantic information. It consists of a predicate and possibly a multitude of arguments. A predicate corresponds to the type of event that occurred, while the arguments

correspond to the entities playing a role in this event. Here, different argument types correspond to different roles. For example, Arg0 is the argument referring to the active entity that initializes the corresponding event. 17, 49, 51, 55, 76, 118, 121, 124, 130, 133

Propbank

is a corpus of financial news that is annotated with semantic roles. It is the standard corpus used to train and evaluate machine learning algorithms to perform semantic role labeling. 12, 18, 80, 108, 113, 118, 121, 131, 148

PubMed

is a "free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH). It comprises over 24 million citations for biomedical literature from MEDLINE, life science journals, and online books" (Pubmed, 2014). 1, 14, 17, 153

Relation

is a description of how two entities are connected. In information extraction relations between entities are often extracted as triples containing of the two connected entities and the type of the relation. 4, 12, 14, 15, 36, 49, 59

Semantic Role Labeling

is the task of assigning semantic roles to parts of sentences with respect to the corresponding predicates. The most important semantic roles correspond to answers to the question "Who did what to whom, when, where, how and why?". 12, 15, 113

Semantics

is the part of linguistics that is concerned with the meaning of utterances. It tries to relate words to entities in the real world and to identify the relations between these entities. Typical semantic analyzers are semantic role labelers. 23, 27, 29, 117, 122, 143, 146, 159

Senna

is a deep-neural-net-based natural language processing tool that - among other NLP tasks - performs state-of-the-art sematnic role labeling. 51, 52, 80, 108, 109, 121, 131, 144

Stemming

is the linguistc procedure to reduce a word to its word stem which is usually used to normalize differently inflected forms of the same word. E.g. "plays", "played" and "playing" would all be reduced to their stem "play". 12, 42

Syntax

is the part of linguistics that is concerned with the rule systems underlying the grammar and more generally the formal structure of a language. Typical syntactic analyzers are constituency parsers based on context-free grammars. 23, 27, 113, 117, 122, 143, 146, 159

Text Mining

is the "analysis of unstructured texts with the goal of uncovering new, previously unknown information" (Hearst, 1999). 4, 8, 152

Tokenization

is the linguistc procedure to split a sentence into tokens. Tokens are usually words. Depending on the respective task, however, words can be split into multiple tokens or multiple words can be combined into one token. 34, 76, 115, 124, 127, 136, 139

Bibliography

Agarwal, S. and Yu, H. (2010). Biomedical negation scope detection with conditional random fields. *Journal of the American Medical Informatics Association*, 17(6):696–701.

Ahn, A. C., Tewari, M., Poon, C.-S. S., and Phillips, R. S. (2006). The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS medicine*, 3(6).

Al-Mubaid, H. and Gungu, S. (2012). A learning-based approach for biomedical word sense disambiguation. *TheScientificWorldJournal*, 2012:949247+.

Al-Shahrour, F., Minguez, P., Tarraga, J., Medina, I., Alloza, E., Montaner, D., and Dopazo, J. (2007). Fatigo+: a functional profiling tool for genomic data. integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res*, 35(Web Server issue).

Alba Juez, L. (2009). *Perspectives on Discourse Analysis: Theory and Practice*. Cambridge Scholars Publishing, Newcastle upon Tyne.

Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F. L., Hakenberg, J., Khelif, K., Schröder, M., and Wächter, T. (2009). Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC Bioinformatics*.

Allan, J. T., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., and Amstutz, P. (2005). Taking topic detection from evaluation to practice. *Proceedings of the 38th annual Hawaii International Conference on System Sciences*.

Ananiadou, S., Kell, D. B., and Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12):571 – 579.

Ananiadou, S., Pyysalo, S., Tsujii, J., and Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

Arighi, C., Lu, Z., Krallinger, M., Cohen, K., Wilbur, W., Valencia, A., Hirschman, L., and Wu, C. (2011). Overview of the biocreative iii workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Armstrong, C. (2010). Aliases and ambiguity: A case study of gene aliases, and implications for information curation and ai.

Ashburner, M. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29.

Austin, J. L. (1975). *How to do things with words*. Harvard University Press.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Barnickel, T. (2009). *Large Scale Knowledge Extraction from Biomedical Literature Based on Semantic Role Labeling*. PhD thesis, Technische Universität München.

Bavisetty S, Grody WW, Y. S. (2013). Emergence of pediatric rare diseases: Review of present policies and opportunities for improvement. *Rare Diseases*, 1.

Beaufort, R., Roekhaut, S., Cougnon, L.-A., and Fairon, C. (2010). A hybrid rule/model-based finite-state framework for normalizing sms messages. In Hajic, J., Carberry, S., and Clark, S., editors, *ACL*, pages 770–779. The Association for Computer Linguistics.

Bedau, M. A. (2003). *Artificial Life*, pages 197–211. Malden, MA: Blackwell Publishing.

Bengio, Y., Courville, A. C., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Bengtson, E. and Roth, D. (2008). Understanding the value of features for coreference resolution. In *EMNLP*, pages 294–303. ACL.

Berthold, M. R., Cebron, N., Dill, F., Fatta, G. D., Gabriel, T. R., Georg, F., Meinl, T., Ohl, P., Sieb, C., and Wiswedel, B. (2006). Knime: The konstanz information miner. Technical report.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2008). Knime: The konstanz information miner. In Preisach, C., Burkhardt, H., Schmidt-Thieme, L., and Decker, R., editors, *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 319–326. Springer, Berlin Heidelberg.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus linguistics : investigating language structure and use*. Cambridge Univ. Press, Cambridge [u.a.].

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., and Salakoski, T. (2009). Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 10–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Björne, J. and Salakoski, T. (2013). Tees 2.1: Automated annotation scheme learning in the bionlp 2013 shared task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 16–25, Sofia, Bulgaria. Association for Computational Linguistics.

Blohm, P., Frishman, G., Smialowski, P., Goebels, F., Wachinger, B., Ruepp, A., and Frishman, D. (2014). Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research*, 42(Database-Issue):396–400.

Blohm, P. and Meiners, S. (2014). Visualization and exploration of large-scale event extraction results. *Submitted to Journal of Biomedical Semantics*.

Blum-Kulka, S. and Hamo, M. (2011). *Discourse Pragmatics*, pages 143–164. SAGE.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Bonial, C., Babko-Malaya, O., Choi, J. D., Hwang, J., and Palmer, M. (2010). *Propbank Annotation Guidelines*.

Borg, C. (2007). Discovering grammar rules for automatic extraction of definitions. In *Doctoral Consortium at the Eurolan Summer School 2007, Iasi, Romania*, pages 61–68.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., and Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. 40:540–545+.

Bui, Q.-C., Campos, D., van Mulligen, E., and Kors, J. (2013). A fast rule-based approach for biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 104–108, Sofia, Bulgaria. Association for Computational Linguistics.

Califf, M. E. and Mooney, R. J. (1999). Relational learning of pattern-match rules for information extraction. In Hendler, J. and Subramanian, D., editors, *AAAI/IAAI*, pages 328–334. AAAI Press / The MIT Press.

Cao, Y., Liu, F., Simpson, P., Antieau, L. D., Bennett, A. S., Cimino, J. J., Ely, J. W., and Yu, H. (2011). Askhermes: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, 44(2):277–288.

Carlson, A., Betteridge, J., Wang, R. C., Jr., E. R. H., and Mitchell, T. M. (2010a). Coupled semi-supervised learning for information extraction. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010)*.

Carlson, A., Betteridge, J., Wang, R. C., Jr., E. R. H., and Mitchell, T. M. (2010b). Coupled semi-supervised learning for information extraction. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 101–110, New York, NY, USA. ACM.

Carlson, A. and Schafer, C. (2008). Bootstrapping information extraction from semi-structured web pages. In Daelemans, W., Goethals, B., and Morik, K., editors, *ECML/PKDD (1)*, volume 5211 of *Lecture Notes in Computer Science*, pages 195–210. Springer.

Carpenter, B. and Baldwin, B. (2011). *Natural Language Processing with LingPipe 4*. LingPipe Publishing, New York, draft edition.

Carroll, M. W. (2011). Why full open access matters. *PLoS Biol*, 9(11):e1001210.

Castellanos, M. (2004). *HotMiner: Discovering Hot Topics from Dirty Text*, pages 123–158. Springer.

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, New York, NY, USA. ACM.

Ceccato, M., Kiyavitskaya, N., Zeni, N., Mich, L., and Berry, D. M. (2004). Ambiguity identification and measurement in natural language texts.

Centrum Cyfrowe, EDRI, Kennisland, Modern Poland Foundation, and La Quadrature du Net (2013). Failure of "licenses for europe" underlines the need for reform of the eu copyright framework. URL = https://www.laquadrature.net/en/failure-of-licenses-for-europe-underlines-the-need-for-reform-of-the-eu-copyright-framework, Accessed: 19/03/2014.

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In Kuipers, B. and Webber, B. L., editors, *AAAI/IAAI*, pages 598–603. AAAI Press / The MIT Press.

Chen, L., Liu, H., and Friedman, C. (2005). Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21:248–256.

Chen, Y., Cao, H., Mei, Q., Zheng, K., and Xu, H. (2013). Applying active learning to supervised word sense disambiguation in medline. *J Am Med Inform Assoc*.

Chen, Z. and Ji, H. (2011). Collaborative ranking: A case study on entity linking. In *EMNLP*, pages 771–781. ACL.

Chieu, H. L. and Ng, H. T. (2002). A maximum entropy approach to information extraction from semi-structured and free text. In *Eighteenth national conference on Artificial intelligence*, pages 786–791, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control*, 2:137–167.

Chomsky, N. and Pinker, S. (2011). Transcript of the q&a of the keynote panel: The golden age - a look at the original roots of artificial intelligence, cognitive science, and neuroscience. URL=http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html, Accessed March 17th 2014.

Cimiano, P., Hotho, A., and Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Int. Res.*, 24(1):305–339.

Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2011). Convolutional neural network committees for handwritten character classification. In *ICDAR*, pages 1135–1139. IEEE.

Coenen, F. (2011). Data mining: past, present and future. *Knowledge Eng. Review*, 26(1):25–29.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71.

Cohen, K. B. and Hunter, L. (2004). *Natural language processing and systems biology*, pages 147–174. Springer.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011a). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011b). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

COREMINE (2014). Coremine medical - explore connections - build your biomedical mind map. URL = http://www.coremine.com/medical/, Accessed: 07/02/2014.

Coulson, A. S., Glasspool, D. W., Fox, J., and Emery, J. (2001). Rags: A novel approach to computerized genetic risk assessment and decision support from pedigrees. *Methods Inf Med*, 40(4):315–22.

Crick, F. H. C. (1966). *Of Molecules and Man*. University of Washington Press, Washington.

Cruchet, S., Gaudinat, A., Rindflesch, T., and Boyer, C. (2009). What about trust in the question answering world? In *Proceedings of the AMIA Annual Symposium*, pages 1–5, San Francisco, USA.

Csaba, G. (2013). *Context based bioinformatics*. PhD thesis, Ludwig-Maximilians-Universität München.

Culotta, A., Wick, M. L., and McCallum, A. (2007). First-order probabilistic models for coreference resolution. In Sidner, C. L., Schultz, T., Stone, M., and Zhai, C., editors, *HLT-NAACL*, pages 81–88. The Association for Computational Linguistics.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech & Language Processing*, 20(1):30–42.

Dale, E. and Chall, S. E. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, 27(1).

Dash, N. S. (2010). Corpus linguistics: a general introduction. *Presented in the Workshop on Corpus Normalization, Linguistic Data Consortium for the Indian Languages (LDCIL), Central Institute of Indian Languages*, 28.

de Beaugrande, R. and Dressler, W. U. (1981). *Introduction to text linguistics*. Longman.

Deane, P. D. (1988). Polysemy and cognition. *Lingua*, 75(4).

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Delmonte, R. (2005). *Deep and shallow linguistically based parsing: Parameterizing ambiguity in a hybrid parser*, pages 335–374. John Benjamins.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology*, 4(5):P3+.

Dethlefs, N. and Cuayáhuitl, H. (2011). Combining hierarchical reinforcement learning and bayesian networks for natural language generation in situated dialogue. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG '11, pages 110–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dezhkam, R. and Khalili, M. (2013). Automatic ontology construction. *J. Basic Appl. Sci. Res.*, 3(1):94–98.

dotplot (2014). dotplot data analysis modeling. Online.

Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jorgensen, H. L., Cox, I. J., Hansen, L. K., Ingwersen, P., and Winther, O. (2013). Findzebra: A search engine for rare diseases.

D'Souza, J. and Ng, V. (2012). Anaphora resolution in biomedical literature: a hybrid approach. In Ranka, S., Kahveci, T., and Singh, M., editors, *BCB*, pages 113–122. ACM.

Earley, J. (1983). An efficient context-free parsing algorithm. *Commun. ACM*, 26(1):57–61.

Erten, S., Bebek, G., Ewing, R. M., and Koyutuerk, M. (2011). Dada: Degree-aware algorithms for network-based disease gene prioritization. *BioData Mining*, 4:19.

Escudero, G., Marquez, L., and Rigau, G. (2000). Naive bayes and exemplar-based approaches to word sense disambiguation revisited. In *In Proceedings of the 14th European Conference on Artificial Intelligence*, pages 421–425.

Eurodis (2005). Eurodis - european organization for rare diseases. rare diseases: understanding this public health priority. URL = http://www.eurordis.org/IMG/pdf/princeps_document-EN.pdf, visited Jan 13th 2014.

Evans, J. and Rzhetsky, A. (2010). Machine Science. *Science*, 329(5990):399–400.

Evans, J. A. and Rzhetsky, A. (2011). Advancing Science through Mining Libraries, Ontologies, and Communities. *Journal of Biological Chemistry*, 286(27):23659–23666.

Evi (2014). Evi. URL = http://evi.com/, Accessed: 12/02/2014.

Fahmi, I. and Bouma, G. (2006). Learning to identify definitions using syntactic features. In *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*.

Felizardo, K. R., Nakagawa, E. Y., Feitosa, D., Minghim, R., and Maldonado, J. C. (2010). An approach based on visual text mining to support categorization and classification in the systematic mapping. In *Proceedings of the 14th international conference on Evaluation and Assessment in Software Engineering*, EASE'10, pages 34–43, Swinton, UK, UK. British Computer Society.

Ferrucci, D. A. (2011). Ibm's watson/deepqa. *SIGARCH Comput. Archit. News*, 39(3):–.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221 – 233.

Fortuna, B., Grobelnik, M., and Mladenic, D. (2005). Visualization of text document corpus. *Informatica (Slovenia)*, 29(4):497–504.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database-Issue):808–815.

Francis, W. N. and Kucera, H. (1979). Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Fundel, K. and Zimmer, R. (2006). Gene and protein nomenclature in public databases. *BMC Bioinformatics*, 7:372.

Gardent, C. and Kallmeyer, L. (2003). Semantic construction in feature-based tag. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 123–130, Stroudsburg, PA, USA. Association for Computational Linguistics.

Garla, V. and Brandt, C. (2012). Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pages 22–22. IEEE.

Gasperin, C. and Briscoe, T. (2008). Statistical anaphora resolution in biomedical texts. In Scott, D. and Uszkoreit, H., editors, *COLING*, pages 257–264.

Gaudio, R. D. and Branco, A. (2007). Automatic extraction of definitions in portuguese: A rule-based approach. In *EPIA Workshops'07*, pages 659–670.

Gerner, M., Nenadic, G., and Bergman, C. M. (2010). Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11:85.

Gerner, M., Sarafraz, F., Bergman, C. M., and Nenadic, G. (2012). BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, 28(16):2154–2161.

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, 1(1):3–55.

Gobeill, J., Patsche, E., Theodoro, D., Veuthey, A.-L., Lovis, C., and Ruch, P. (2009). Question answering for biology and medicine. In *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, page 1–5. IEEE, IEEE.

Gooch, P. (2012). Badrex: In situ expansion and coreference of biomedical abbreviations using dynamic regular expressions. *CoRR*, abs/1206.4522.

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the blast2go suite. *Nucleic Acids Res*, 36(10):3420–35.

Graber, M. L. and Mathew, A. (2008). Performance of a web-based clinical diagnosis support system for internists. *J Gen Intern Med*, 23 Suppl 1:37–40.

Graesser, A. C., Mcnamara, D. S., Louwerse, M. M., Cai, Z., Dempsey, K., Floyd, Y., Mccarthy, P., Ozuru, Y., Petrowski, M., Pillarisetti, S., Reese, M., Rowe, M., Sayroo, J., Sumara, K., and Correspondence, F. Y. (2004). Coh-metrix: Analysis of text on cohesion and language. In *Topics in Cognitive Science*, page 27.

Greek-Winald, C., Gustafsson, B., and Högvik, L. (2010). The swedish rare disease information database and the swedish information centre for rare diseases. *Orphanet Journal of Rare Diseases*, O26(5(Suppl 1)).

Green, G. M. (1989). *Pragmatics and Natural Language Understanding*. Erlbaum, Hillsdale, NJ.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, 339(jul20_3):b2680++.

Greiner, H. M., Horn, P. S., Holland, K., Collins, J., Hershey, A. D., and Glauser, T. A. (2013). mRNA blood expression patterns in new-onset idiopathic pediatric epilepsy. *Epilepsia*, 54(2):272–279.

Grice, H. (1957). Meaning. *The Philosophical Review*, 66:377–88.

Grice, P. (1981). Utterer's Meaning, Sentence-Meaning, and Word-Meaning. In *Studies in the Way of Words*, pages 117–137. Harvard University Press, Cambride, MA.

Gries, S. T. (2010). Corpus linguistics and theoretical linguistics A love, hate relationship? Not necessarily. *International Journal of Corpus Linguistics*, 15:327–343.

Grover, C. and Tobin, R. (2006). Rule-based chunking and reusability. In *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC*.

Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2013). Evex in st'13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 26–34, Sofia, Bulgaria. Association for Computational Linguistics.

Hakenberg, J., Gerner, M., Haeussler, M., Solt, I., Plake, C., Schroeder, M., Gonzalez, G., Nenadic, G., and Bergman, C. M. (2011). The gnat library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Han, X., Sun, L., and Zhao, J. (2011). Collective entity linking in web text: a graph-based method. In Ma, W.-Y., Nie, J.-Y., Baeza-Yates, R. A., Chua, T.-S., and Croft, W. B., editors, *SIGIR*, pages 765–774. ACM.

Hargreaves, I. and Office, U. I. P. (2011). *Digital Opportunity: A Review of Intellectual Property and Growth ; an Independent Report*. Intellectual Property Office.

Hatzigaidas, A., Papastergiou, A., Tryfon, G., and Zaharias, Z. (2004). *Topic Map Existing Tools: A Brief Review*, pages 16–18.

Hatzivassiloglou, V., Duboue, P. A., and Rzhetsky, A. (2001). Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach.

Havasi, C., Speer, R., and Alonso, J. (2007). Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Hazman, M., El-Beltagy, S. R., and Rafea, A. (2011). A survey of ontology learning approaches. *International Journal of Computer Applications*, 22(8):36–43. Published by Foundation of Computer Science.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 3–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heer, J. (2005). Prefuse: a toolkit for interactive information visualization. In *In CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM Press.

Herzog, M. (2014). Generation and consolidation of a controlled vocabulary for the development of an integrated knowledge base for rare diseases.

HHS (1989). United states department of health and human services. report of the national commission on orphan diseases.

Hirschman, L., Yeh, A. S., Blaschke, C., and Valencia, A. (2005a). Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(S-1).

Hirschman, L., Yeh, A. S., Blaschke, C., and Valencia, A. (2005b). Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(S-1).

Hirschmann, H. (2014). Stts-tagset gemäss tiger annotationsschema. URL = http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger/view. Accessed: 10/03/2014.

Hobbs, J. R. and Riloff, E. (2010). Information extraction. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Hoffmann, R. and Valencia, A. (2005). Implementing the ihop concept for navigation of biomedical literature. In *ECCB/JBI*, page 258.

Hosur, R., Peng, J., Vinayagam, A., Stelzl, U., Xu, J., Perrimon, N., Bienkowska, J., and Berger, B. (2012). A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. *Genome biology*, 13(8):R76+.

Huang, M., Liu, J., and Zhu, X. (2011). Genetukit: a software for document-level gene normalization. *Bioinformatics*, 27(7):1032–1033.

Hur, J., Schuyler, A. D., States, D. J., and Feldman, E. L. (2009). Sciminer: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics.*, 25(16):838–840.

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. (2008). A mixed method lemmatization algorithm using a hierarchy of linguistic identities (holi). In *Proceedings of the 6th international conference on Advances in Natural Language Processing*, GoTAL '08, pages 205–216, Berlin, Heidelberg. Springer-Verlag.

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart (2013). Stts tag table (1995/1999). URL = http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html. Accessed:

10/03/2014.

Jeske, T. (2013). Functional analysis of gene lists by visualization and network analysis of text mining results within gene ontology classes. Bachelor's thesis, Ludwig-Maximilians Universität München, Technische Universität München.

Jiang, M., Jensen, E., Beitzel, S., and Argamon, S. (2004). Choosing the right bigrams for information retrieval. In *In Proceeding of the Meeting of the International Federation of Classification Societies*.

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R. M., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2012). David-ws: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806.

Kanehisa, M. and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kang, N., Afzal, Z., Singh, B., van Mulligen, E. M., and Kors, J. A. (2012). Using an ensemble system to improve concept extraction from clinical records. *Journal of Biomedical Informatics*, 45(3):423–428.

Kaufmann, M. and Kalita, J. (2010). Syntactic normalization of Twitter messages. In *Proceedings of the 8^{th} International Conference on Natural Language Processing (ICON 2010)*, Chennai, India. Macmillan India.

Kemper, B., Matsuzaki, T., Matsuoka, Y., Tsuruoka, Y., Kitano, H., Ananiadou, S., and ichi Tsujii, J. (2010). Pathtext: a text mining integrator for biological pathway visualizations. *Bioinformatics [ISMB]*, 26(12):374–381.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Röchert, B., Orchard, S. E., and Hermjakob, H. (2012). The intact molecular interaction database in 2012. *Nucleic Acids Research*, 40(Database-Issue):841–846.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., Bader, G. D., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J. J., Moore, S., Ceol, A., Chatr-Aryamontri, A., Oesterheld, M., Stümpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M. E., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., and Hermjakob, H. (2007). Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC biology*, 5(1):44+.

Kikui, G.-i. (1992). Feature structure based semantic head driven generation. In *COLING 1992 Volume 1: The 15th International Conference on Computational Linguistics*.

Kim, J.-D., Ohta, T., Tateisi, Y., and ichi Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*, pages 180–182.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *In Language Resources and Evaluation*.

Kitano, H. (2002). Systems biology: A brief overview.

Klavans, J. L. and Muresan, S. (2001). *Evaluation of DEFINDER: a system to mine definitions from consumer-oriented medical text*, pages 201–202. ACM Press.

KNIME development team (2013). Knime labs. URL = http://tech.knime.org/knime-labs. Accessed: 21/11/2013.

Kobyliński, L. and Przepiórkowski, A. (2008). Definition extraction with balanced random forests. In *Proceedings of the 6th international conference on Advances in Natural Language Processing*, GoTAL '08, pages 237–247, Berlin, Heidelberg. Springer-Verlag.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics*, (4):457–64.

Konev, B., Schmidt, R. A., and Schulz, S. (2010). Special issue on practical aspects of automated reasoning. *AI Commun.*, 23(2-3):67–68.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J. D., Bourne, P. E., and Berman, H. M. (2006). The rcsb pdb information portal for structural genomics. *Nucleic Acids Research*, 34(Database-Issue):302–305.

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., and Valencia, A. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biology*, 9(Suppl 2):S1+.

Kuperman, G., Gardner, R., and Pryor, T. (1991). *Help: A Dynamic Hospital Information System*. Computers and medicine. Springer-Verlag.

Ladyman, J., Lambert, J., and Wiesner, K. (2011). What is a complex system?

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Landauer, T., Foltz, P., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25:259–284.

Larson, R. R. (2012). Information retrieval systems. In Bates, M. J., editor, *Understanding Information Retrieval Systems - Management, Types, and Standards*. CRC Press.

Latreche, S. (2011). Sprachdialogsysteme mit chunk-parsing. Magisterarbeit, Universität Duisburg-Essen.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., and Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. In *ICML*. icml.cc / Omnipress.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551.

Leser, U. and Hakenberg, J. (2005). What makes a gene name? named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–369.

Levinson, S. C. (1983). *Pragmatics (Cambridge Textbooks in Linguistics)*. Cambridge University Press.

Lewis, J. (2000). The logic theorist and its children: Ai in action. URL = http://www.cs.swarthmore.edu/~eroberts/cs91/projects/ethics-of-ai/sec1_2.html. Accessed: 28/10/2013.

Li, L., Wang, Y., and Huang, D. (2013). Improving feature-based biomedical event extraction system by integrating argument information. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 109–115, Sofia, Bulgaria. Association for Computational Linguistics.

Liu, H., Keselj, V., Blouin, C., and Verspoor, K. (2012). Subgraph matching-based literature mining for biomedical relations and events. In *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, volume FS-12-05 of *AAAI Technical Report*. AAAI.

Liu, H., Verspoor, K., Comeau, D. C., MacKinlay, A., and Wilbur, W. J. (2013a). Generalizing an approximate subgraph matching-based system to extract events in molecular biology and cancer genetics. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 76–85, Sofia, Bulgaria. Association for Computational Linguistics.

Liu, X., Bordes, A., and Grandvalet, Y. (2013b). Biomedical event extraction by multi-class classification of pairs of text entities. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 45–49, Sofia, Bulgaria. Association for Computational Linguistics.

London, S. (1998). Dxplain: a web-based diagnostic decision support system for medical students. *Med Ref Serv Q*, 17(2):17–28.

Lopez, V., Uren, V., Sabou, M., and Motta, E. (2011). Is question answering fit for the semantic web?: A survey. *Semant. web*, 2(2):125–155.

Lussier, Y. A., Borlawsky, T., Rappaport, D., Liu, Y., and Friedman, C. (2006). Phenogo: Assigning phenotypic context to gene ontology annotations with natural language processing. In Altman, R. B., Murray, T., Klein, T. E., Dunker, A. K., and Hunter, L., editors, *Pacific Symposium on Biocomputing*, pages 64–75. World Scientific.

Lutzenberger, T. (2014). Context model based co-reference resolution in biomedical literature. Not submitted.

MacKinlay, A., Martinez, D., Jimeno Yepes, A., Liu, H., Wilbur, W. J., and Verspoor, K. (2013). Extracting biomedical events and modifications using subgraph matching with noisy training data. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 35–44, Sofia, Bulgaria. Association for Computational Linguistics.

Maedche, A., Neumann, G., and Staab, S. (2003). Bootstrapping an ontology-based information extraction system.

Malheiros, V., Hoehn, E. N., Pinho, R., Mendonca, M. G., and Maldonado, J. C. (2007). A visual text mining approach for systematic reviews. In *ESEM*, pages 245–254. IEEE Computer Society.

Markoff, J. (2012). How many computers to identify a cat? 16,000. Published on nytimes.com, URL = "http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-

machine-learning.html?pagewanted=all&_r=0", Accessed: Mar 12th 2014.

Marsh, G. E. (2009). The demystification of emergent behavior. *East*, page 9.

Martin, W. A., Church, K. W., and Patil, R. S. (1987). Preliminary analysis of the breadth-first parson algorithm: theoretical and experimental results. In Bolc, L., editor, *Natural Language Parsing Systems*. Springer.

Martinez, A. R. (2012). Part-of-speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(1):107–113.

Matsukawa, J., Snodgrass, J. G., and Doniger, G. M. (2005). Conceptual versus perceptual priming in incomplete picture identification. *J Psycholinguist Res*, 34(6):515–40.

Mazzocchi, F. (2008). Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory. *EMBO reports*, 9(1):10–14.

McCallum, A. and Wellner, B. (2005). Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Processing Systems 17*, pages 905–912. MIT Press.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *IJCAI-95*, pages 1050–1055, Montreal, Canada.

McDonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis the influence of context on judgements of semantic similarity. In *23rd Annual Conference of the Cognitive Science Society*.

McEnery, T. and Wilson, A. (2005). *Corpus linguistics : an introduction*. Edinburgh Univ. Press, Edinburgh.

MedDRA (2014). Meddra - medical dictionary for regulatory activities. URL = http://www.meddra.org/, visited Mar 19th 2014.

Medina, I., Carbonell, J., Pulido, L., Madeira, S. C., G"otz, S., Conesa, A., Tarraga, J., Pascual-Montano, A. D., Nogales-Cadenas, R., Santoyo, J., Garcia, F., Marba, M., Montaner, D., and Dopazo, J. (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Research*, 38(Web-Server-Issue):210–213.

Mei, Q. (2009). *Contextual text mining*. PhD thesis, Graduate College of the University of Illinois at Urbana-Champaign.

Meier, J.-M. (2014). Erstellung einer ressource für ein diagnoseunterstützungssystem für seltene erkrankungen.

Metz, C. (2013). Facebook's 'deep learning' guru reveals the future of ai. Published on wired.com, URL = "http://www.wired.com/wiredenterprise/2013/12/facebook-yann-lecun-qa/", Accessed: Mar 12th 2014.

Meyer, D. E. and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

Meyer, D. E., Schvaneveldt, R. W., and Ruddy, M. G. (1975). *Loci of contextual effects on visual word-recognition*, chapter 8, pages 98–118. London: Academic Press.

Middleton, B., Shwe, M., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., and Cooper, G. (1990). Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base - II. Evaluation of Diagnostic Performance. *Medicine*, 30:241–255.

Miliaraki, S. and Androutsopoulos, I. (2004). Learning to identify single-snippet answers to definition questions. In *Proceedings of COLING-2004*, pages 1360–1366.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Miner, G., Elder, J., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Elsevier Science.

Minsky, M. and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

Mooers, C. N. (1950). Information retrieval viewed as temporal signalling. In *Proceedings International Congress of Mathematicians*, volume 1, page 572.

Morante, R. and Sporleder, C., editors (2010). *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, Sweden. University of Antwerp.

Morgan, A. A. (2005). Linking text mentions to biological identifiers. Invited talk, February 4, 2005, Ontario Centre for Genomic Computing, Text Mining Tools for Bioinformaticians and Biologists Workshop.

Morris, C. W. (1938). *Foundations of the Theory of Signs*. University of Chicago Press, Chicago, IL, 1st edition.

Morton, T. S. (2000). Coreference for nlp applications. In *ACL*. ACL.

Murata, M., Utiyama, M., Uchimoto, K., Ma, Q., and Isahara, H. (2001). Japanese word sense disambiguation using the simple bayes and support vector machine methods.

Murphy, M. L. (2010). *Lexical Meaning*. Cambridge University Press.

Murray-Rust, P., Molloy, J., and Cabell, D. (2012). Open Content Mining. *Conference for the Fellows of OpenForum Academy - 24th September 2012 Brussels*, pages 52–58.

Muslea, I., Minton, S., and Knoblock, C. (1998). Wrapper induction for semistructured web-based information sources. In *Proceedings of the Conference on Automated Learning and Discovery (CONALD), 1998*.

Myers, J. D. (1987). The background of internist i and qmr. In Blum, B. I., editor, *History of Medical Informatics*, pages 195–197. ACM.

Nahm, U. Y. (2004). *Text Mining with Information Extraction*. PhD thesis, Department of Computer Sciences, University of Texas at Austin.

Nahm, U. Y., Bilenko, M., and Mooney, R. J. (2002). Two approaches to handling noisy variation in text mining. In *Papers from the Nineteenth International Conference on Machine Learning (ICML-2002) Workshop on Text Learning*, pages 18–27, Sydney, Australia.

Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-J., Ohta, T., Pyysalo, S., and Zweigenbaum, P. (2013). Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria. Association for Computational Linguistics.

Nenkova, A. and McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Newman, S. A. (2003). The fall and rise of systems biology. *GeneWatch*, 16(4):8–12.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Ng, V. (2008). Unsupervised models for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 640–649.

Ng, V. and Gardent, C. (2002). Improving machine learning approaches to coreference resolution. In *ACL*, pages 104–111. ACL.

Niemann, H., Sagerer, G., Schröder, S., and Kummert, F. (1990). Ernest: A semantic network system for pattern understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(9):883–905.

Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. (2003). Pathway studio - the analysis and navigation of molecular networks. *Bioinformatics*, 19(16):2155–2157.

Nissim, M. and Markert, K. (2003). Syntactic features and word similarity for supervised metonymy resolution. In Hinrichs, E. W. and Roth, D., editors, *ACL*, pages 56–63. ACL.

Noble, I. (2003). Human genome finally complete. http://news.bbc.co.uk/2/hi/science/nature/2940601.stm. Accessed: 13/03/2012.

Norvig, P. (2011). On chomsky and the two cultures of statistical learning. URL=http://norvig.com/chomsky.html, Accessed March 17th 2014.

Ohta, T., Tateisi, Y., and Kim, J. (2002). The genia corpus: An annotated research abstract corpus in molecular biology domain. In *the Human Language Technology Conference*.

OMIM (2014). Online mendelian inheritance in man, omim. mckusick-nathans institute of genetic medicine, johns hopkins university (baltimore, md). URL = http://omim.org/, visited Jan 14th 2014.

ORDR (2014). Office of rare diseases research, national center for advancing translational sciences (ncats). URL = http://rarediseases.info.nih.gov/, visited Jan 14th 2014.

Orphanet (2014). Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at http://www.orpha.net Accessed Jan 14th 2014.

Osborne, G., Cable, V., and Hunt, J. (2011). The government response to the hargreaves review of intellectual property and growth. URL = http://www.ipo.gov.uk/ipresponse-full.pdf, Accessed: 18/03/2014.

Pagon RA, Adam MP, Bird TD, et al. (2014). Genereviews. Seattle (WA): University of Washington, Seattle; 1993-2014. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1116/, visited Jan 14th 2014.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1).

Palmer, M., Gildea, D., and Xue, N. (2010). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Paul, D. B. and Baker, J. M. (1992). The design for the wall street journal-based csr corpus. In *ICSLP*. ISCA.

Pecheux, M. (1995). *Automatic Discourse Analysis*. Rodopi Bv Editions.

Pedersen, T. (2006). Unsupervised Corpus-Based Methods for WSD. In *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, pages 133–166. Springer.

Pedersen, T., Banerjee, S., and Patwardhan, S. (2003). Maximizing semantic relatedness to perform word sense disambiguation.

Pham, X. Q., Le, M. Q., and Ho, B. Q. (2013). A hybrid approach for biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 121–124, Sofia, Bulgaria. Association for Computational Linguistics.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, 92:377–410.

Planas-Iglesias, J., Bonet, J., García-García, J., Marín-López, M. A., Feliu, E., and Oliva, B. (2013). Understanding protein-protein interactions using local structural features. *Journal of molecular biology*, 425(7):1210–1224.

PMC (2014). U.s. national library of medicine national center for biotechnology information: Pubmed central. URL = http://www.ncbi.nlm.nih.gov/pubmed, Accessed: 04/02/2014.

Preiss, J. and Stevenson, M. (2013). Dale: A word sense disambiguation system for biomedical documents trained using automatically labeled examples. In *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Przepiórkowski, A., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kuboň, V., and Wójtowicz, B. (2007). Towards the automatic extraction of definitions in slavic. In *Proceedings ofo the BSNLP workshop at ACL 2007*.

Pubmed (2014). U.s. national library of medicine national center for biotechnology information: Pubmed. URL = http://www.ncbi.nlm.nih.gov/pubmed, Accessed: 04/02/2014.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. (2007a). Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50.

Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J., and Salakoski, T. (2007b). BioInfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):50+.

R., H.-N. (1994). Context vectors: general purpose approximate meaning representations self-organized from raw data. *Computational Intelligence: Imitating Life, IEEE Press*, pages 43–56.

Ramshaw, L. A. and Marcus, M. P. (1995). Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040.

Rapid-I (2014). RapidMiner. Online.

Reisberg, D. (2005). *Cognition: Exploring the Science of the Mind*. WW Norton, New York, 3 edition.

Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Clematide, S., Hess, M., von Allmen, J.-M., Parisot, P., Romacker, M., and Vachon, T. (2008). Ontogene in biocreative ii. *Genome Biology*, 9(Suppl 2):S13.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–615.

Roller, R. and Stevenson, M. (2013). Identification of genia events using multiple classifiers. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 125–129, Sofia, Bulgaria. Association for Computational Linguistics.

Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G., and Gurcan, M. N. (2013). Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of pathology informatics*, 4.

Royer, L., Reimann, M., Stewart, A. F., and Schröder, M. (2012). Network Compression as a Quality Measure for Protein Interaction Networks. *PLoS ONE*, 7(6):e35729+.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

Rumelhart, D., Hintont, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Sagae, K. and ichi Tsujii, J. (2007). Dependency parsing and domain adaptation with lr models and parser ensembles. In *EMNLP-CoNLL*, pages 1044–1050. ACL.

Sagae, K., Miyao, Y., and ichi Tsujii, J. (2007). Hpsg parsing with shallow dependency constraints. In Carroll, J. A., van den Bosch, A., and Zaenen, A., editors, *ACL*. The Association for Computational Linguistics.

Saggion, H. (2004). Identifying definitions in text collections for question answering. lrec. In *LREC 2004*. LREC.

Sahlgren, M. (2005). An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

Sahlgren, M. (2006). *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. PhD thesis, Stockholm University, Stockholm, Sweden.

Salton, G. M., Wong, A. K. C., and Yang, C.-S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.

Sanchez, D. and Moreno, A. (2004). Creating ontologies from web documents. In *Recent Advances in Artificial Intelligence Research and Development*, volume 113, pages 11–18. IOS Press.

Sanchez-Graillet, O. and Poesio, M. (2007). Negation of protein-protein interactions: analysis and extraction. In *ISMB/ECCB (Supplement of Bioinformatics)*, pages 424–432.

Sang, E. F. T. K. and Buchholz, S. (2000). Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning, September*, pages 13–14.

Sappelt, F. (2013). Complex disease identification by shallow semantic fact extraction (ssfe) in german patient records. Master's thesis, Ludwig-Maximilians Universität München, Technische Universität München.

Sarafraz, F. and Nenadic, G. (2010). Identification of negated regulation events in the literature: exploring the feature space. In Collier, N., Hahn, U., Rebholz-Schuhmann, D., Rinaldi, F., and Pyysalo, S., editors, *Semantic Mining in Biomedicine*, volume 714 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Scherf, M., Epple, A., and Werner, T. (2005). The next generation of literature analysis: Integration of genomic analysis into text mining. *Briefings in Bioinformatics*, 6(3):287–297.

Schneider, G., Clematide, S., Ellendorff, T., Tuggener, D., Rinaldi, F., and Grigonyte, G. (2013). Uzh in bionlp 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 116–120, Sofia, Bulgaria. Association for Computational Linguistics.

Schuemie, M. J., Weeber, M., Schijvenaars, B. J. A., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B., and Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications.

Schvaneveldt, R. W. and Meyer, D. E. (1973). Retrieval and comparison processes in semantic memory. In Kornblum, S., editor, *Attention and Performance IV*. Academic Press, New York.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.

Sebastiani, F. (2005). *Text categorization*, pages 109–129. Advances in Management Information. WIT Press.

Segal, M. (2004). Systems and methods for diagnosing medical conditions. US Patent 6,754,655.

Shaoul, C. and Westbury, C. (2010). The westbury lab wikipedia corpus. Website. AB: University of Alberta (downloaded from http://www.psych.ualberta.ca/ westburylab/downloads/westburylab.wikicorp.download.html).

Shapiro, S. C. and Rapaport, W. J. (1986). Sneps considered as a fully intensional propositional semantic network. In Kehler, T., editor, *AAAI*, pages 278–283. Morgan Kaufmann.

Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Research and Development in Information Retrieval*, pages 21–29.

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D., and Ruepp, A. (2010). The negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Research*, 38(Database-Issue):540–544.

Smith, C., Goldsmith, C. A., and Eppig, J. (2004). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7+.

Smith, D., , Smith, D., and Lopez, M. (1997). Information extraction for semi-structured documents. In *In Proceedings of the Workshop on Management of Semistructured Data*.

Smuts, J. C. (1926). *Holism and evolution,*. New York,The Macmillan company,.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Soscia, S. J., Kirby, J. E., Washicosky, K. J., Tucker, S. M., Ingelsson, M., Hyman, B., Burton, M. A., Goldstein, L. E., Duong, S., Tanzi, R. E., and Moir, R. D. (2010). The Alzheimer's disease-associated amyloid beta-protein is an antimicrobial peptide. *PLoS ONE*, 5(3).

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. (2011). The german traffic sign recognition benchmark: A multi-class classification competition. In *IJCNN*, pages 1453–1460. IEEE.

Stark, C., Breitkreutz, B.-J. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535–D539.

Storrer, A. and Wellinghoff, S. (2006). Automated detection and annotation of term definitions in german text corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 2373–2376.

Strache, R. (2012). Inference of meta-information from biomedical semantic role labelling. Master's thesis, Ludwig-Maximilians Universität München, Technische Universität München.

Stumpf, M., Thorne, T., de Silva, E., Stewart, R., An, H., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. 105:6959–6964+.

Stylios, C. D., Georgopoulos, V. C., Malandraki, G. A., and Chouliara, S. (2008). Fuzzy cognitive map architectures for medical decision support systems. *Appl. Soft Comput.*, 8(3):1243–1251.

Su, J., Yang, X., Hong, H., Tateisi, Y., and Tsujii, J. (2008). Coreference resolution in biomedical texts: a machine learning approach. In Ashburner, M., Leser, U., and Rebholz-Schuhmann, D., editors, *Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives*, number 08131 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In Bhargava, B. K., Finin, T. W., and Yesha, Y., editors, *CIKM*, pages 67–74. ACM.

Swanson, D. (1986). Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspect. Bio. Med*, 30:7–18.

Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspect Biol Med*, 31(4):526–557.

Swanson, D. R. and Smalheiser, N. R. (1994). Assessing a gap in the biomedical literature - magnesium-deficiency and neurologic disease. *Neuroscience Research Communications*, 15(1):1–9.

Szarvas, G., Vincze, V., Farkas, R., and Csirik, J. (2008). The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio. Association for Computational Linguistics.

Tamames, J. and de Lorenzo, V. (2010). Envmine: A text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics*, 11:294.

Tari, L., Anwar, S., Liang, S., Cai, J., and Baral, C. (2010). Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, 26(18).

Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: An overview.

The Apache Software Foundation (2010). Opennlp website. URL = http://opennlp.apache.org, visited Dec 19th 2013.

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., and Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, 33(5):1544–1552.

Titscher, S., Meyer, M., Wodak, R., and Vetter, E. (2000). *Methods of Text and Discourse Analysis*. Sage, London.

Tomasello, M. (2000). A usage-based approach to child language acquisition. *Proceedings of the Twenty-Sixth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Aspect (2000)*, pages 305–319.

Tomita, M. (1985). An efficient context-free parsing algorithm for natural languages. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'85, pages 756–764, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Trewavas, A. (2006). A brief history of systems biology. "Every object that biology studies is a system of systems." Francois Jacob (1974). 18(10):2420–30+.

Trigui, O., Belguith, L. H., and Rosso, P. (2010). An automatic definition extraction in arabic language. In *Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems*, NLDB'10, pages 240–247, Berlin, Heidelberg. Springer-Verlag.

Tsai, R. T.-H., Chou, W.-C., Su, Y.-S., Lin, Y.-C., Sung, C.-L., Dai, H.-J., Yeh, I. T.-H., Ku, W., Sung, T.-Y., and Hsu, W.-L. (2007). Biosmile: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, 8.

Tuason, O., Chen, L., Liu, H., Blake, J. A., and Friedman, C. (2004). Biological nomenclatures: A source of lexical knowledge and ambiguity. In Altman, R. B., Dunker, A. K., Hunter, L., Jung, T. A., and Klein, T. E., editors, *Pacific Symposium on Biocomputing*, pages 238–249. World Scientific.

Turian, J., Ratinov, L., and Bengio, Y. (2010a). Word representations for nlp. Website. downloaded from http://metaoptimize.com/projects/wordreprs/.

Turian, J., Ratinov, L.-A., and Bengio, Y. (2010b). Word Representations: A Simple and General Method for Semi-Supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)*, 37:141–188.

University, S. (2011). Corenlp.

Uryupina, O. (2010). Corry: A system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 100–103, Stroudsburg, PA, USA. Association for Computational Linguistics.

Valente, G. T., Acencio, M. L., Martins, C., and Lemke, N. (2013). The development of a universal in silico predictor of protein-protein interactions. *PLoS One*, 8(5):e65587.

van Dijk, T. A. (1977). *Text and Context: Exploration in the Semantics and Pragmatics of Discourse*.

Van Landeghem, S., Björne, J., Wei, C.-H., Hakala, K., Pyysalo, S., Ananiadou, S., Kao, H.-Y., Lu, Z., Salakoski, T., Van de Peer, Y., and Ginter, F. (2013). Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, 8(4).

Van Noorden, R. (2013). Tensions grow as data-mining discussions fall apart. *Nature*, 498(7452):14–15.

Viola, P. A. and Narasimhan, M. (2005). Learning to extract information from semi-structured text using a discriminative context free grammar. In Baeza-Yates, R. A., Ziviani, N., Marchionini, G., Moffat, A., and Tait, J., editors, *SIGIR*, pages 330–337. ACM.

Vita, R., Vaughan, K., Zarebski, L., Salimi, N., Fleri, W., Grey, H., Sathiamurthy, M., Mokili, J., Bui, H.-H., Bourne, P. E., Ponomarenko, J. V., de Castro Jr., R., Chan, R. K., Sidney, J., Wilson, S. S., Stewart, S., Way, S., Peters, B., and Sette, A. (2006). Curation of complex, context-dependent immunological data. *BMC Bioinformatics*, 7:341.

Vogt, P. (2005). Editorial: Language acquisition and evolution. *Adaptive Behavior*, 13(4):325 – 346.

von Mering, C., Huynen, M. A., Jäggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). String: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database-Issue):433–437.

Voutilainen, A. (2004). Part-of-Speech Tagging. In Mitkov, R., editor, *The Oxford handbook of computational linguistics*, chapter 11, pages 219–232. Oxford University Press, New York.

Wachinger, B. N. X. (2013). *Next Generation Knowledge Extraction from Biomedical Literature with Semantic Big Data Approaches*. PhD thesis, Ludwig-Maximilians Universität München, Technische Universität München.

Wald, R., Khoshgoftaar, T. M., Napolitano, A., and Sumner, C. (2012). Using twitter content to predict psychopathy. In *ICMLA (2)*, pages 394–401. IEEE.

Waldvogel, B. (2013). liblinear-java - java port of the original c++ sources. Website. URL = http://liblinear.bwaldvogel.de/, visited Dec 22nd 2013.

Walter, S. and Pinkal, M. (2006). Automatic extraction of definitions from german court decisions. In *Proceedings of the workshop on information extraction beyond the document*, pages 20–28.

Wästfelt, M., Fadeel, B., and Henter, J.-I. (2006). A journey of hope: lessons learned from studies on rare diseases and orphan drugs. *J Intern Med*, 260(1):1–10.

Waxman, S., Fu, X., Arunachalam, S., Leddon, E., Geraghty, K., and Song, H.-J. (2013). Are nouns learned before verbs? infants provide insight into a longstanding debate. *Child Dev Perspect*, 7(3).

Wei, Q. and Collier, N. (2011). Towards classifying species in systems biology papers using text mining. *BMC research notes*, 4(1).

Wells, S., Kerr, A., Broad, J., Riddell, T., Kenealy, T., and Jackson, R. (2007). The impact of new zealand cvd risk chart adjustments for family history and ethnicity on eligibility for treatment (predict cvd-5). *N Z Med J*, 120(1261):U2712.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University.

Westerhout, E. (2009). Definition extraction using linguistic and structural features. In *Proceedings of the 1st Workshop on Definition Extraction*, WDE '09, pages 61–67, Stroudsburg, PA, USA. Association for Computational Linguistics.

Widdows, D. and Ferraro, K. (2008). Semantic vectors: a scalable open source package and online technology management application. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 1183–1190.

Wilbanks, J. (2013). Licence restrictions: A fool's errand. *Nature*, 495(7442):440–441.

Winkler, A.-J. (2011). Semantic named entity type disambiguation via bayesian inference. Master's thesis, Ludwig-Maximilians-Universit"at, Technische Universit"at M"unchen, Germany.

World Health Organization (2014). International classification of diseases (icd). URL = http://www.who.int/classifications/icd/en/, visited Mar 19th 2014.

Yan, H., Jiang, Y., Zheng, J., Peng, C., and Li, Q. (2006). A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Syst. Appl.*, 30(2):272–281.

Yatsko, V. (1998). Integrational discourse analysis. URL = http://vetsky.narod2.ru/IDA, visited Jan 16th 2013.

Yearzunis, W. S. (2006). The crm114 discriminator revealed! - or - how i learned to stop worrying and love my automatic monitoring systems.

Yerazunis, W. S. (2004). The Spam-Filtering Accuracy Plateau at 99.9 percent Accuracy and How to Get Past It. In *MIT Spam Conference 2004*.

Zhang, W., Su, J., Tan, C. L., and Wang, W. (2010). Entity linking leveraging automatically generated annotation. In Huang, C.-R. and Jurafsky, D., editors, *COLING*, pages 1290–1298. Tsinghua University Press.

Zheng, J., Chapman, W. W., Crowley, R. S., and Savova, G. K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44(6):1113–1122.

Zweigenbaum, P., Deleger, L., Lavergne, T., Neveol, A., and Bodnari, A. (2013). A supervised abbreviation resolution system for medical text. *CLEF 2013 Evaluation Labs and Workshop Online Working Notes, Valencia, Spain, 2013*.

Publication Record

At the time of submission of this thesis, the following publications (journal articles and conference proceedings) with my participation and an association to the work presented in this thesis have been published, submitted, or were in preparation:

[1] Blohm, P., Wachinger, B., Gross, N., Barnickel, T., and Stümpflen, V. (2014). DefineTHAT – Automated literature-based extraction of definitions for biomedical terms. *In preparation*.

[2] Jeske, T., Blohm, P., and Mewes, H. W. (2014). Functional Analysis of Gene Lists by Visualization and Network Analysis of Text Mining Results within Gene Ontology Classes. *In preparation*.

[3] Blohm, P. and Meiners, S. (2014). Visualization and exploration of large-scale event extraction results. *Submitted to Journal of Biomedical Semantics*.

[4] Blohm, P., Frishman, G., Smialowski, P., Goebels, F., Wachinger, B., Ruepp, A., and Frishman, D. (2014). Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research*, 42, pages 396-400.

Declaration / Erklärung

Ich erkläre an Eides statt, dass ich die bei der promotionsführenden Einrichtung Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Promotionsprüfung vorgelegte Arbeit mit dem Titel

Supersemantics for Knowledge Extraction

am Institut für Bioinformatik und Systembiologie des Helmholtz Zentrum München unter der Anleitung und Betreuung durch *Prof. Dr. Hans-Werner Mewes* ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß § 6 Abs. 6 und 7 Satz 2 angegebenen Hilfsmittel benutzt habe.

- ☒ Ich habe keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen und Betreuer für die Anfertigung von Dissertationen sucht, oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt.
- ☒ Ich habe die Dissertation in dieser oder ähnlicher Form in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt.
- □ Die vollständige Dissertation wurde in ..... veröffentlicht. Die promotionsführende Einrichtung ..... hat der Vorveröffentlichung zugestimmt.
- ☒ Ich habe den angestrebten Doktorgrad noch nicht erworben und bin nicht in einem früheren Promotionsverfahren für den angestrebten Doktorgrad endgültig gescheitert.
- □ Ich habe bereits am ..... bei der Fakultät für ..... der Hochschule ..... unter Vorlage einer Dissertation mit dem Thema ..... die Zulassung zur Promotion beantragt mit dem Ergebnis: ..... .

Die öffentlich zugängliche Promotionsordnung der Technischen Universität München ist mir bekannt, insbesondere habe ich die Bedeutung von §28 (Nichtigkeit der Promotion) und §29 (Entzug des Doktorgrades) zur Kenntnis genommen. Ich bin mir der Konsequenzen einer falschen Eidesstattlichen Erklärung bewusst.

Mit der Aufnahme meiner personenbezogenen Daten in die Alumni-Datei bei der Technischen Universität München bin ich

- ☒ einverstanden
- □ nicht einverstanden

München, den ....

Philipp Blohm