A STOCHASTIC GRAMMAR FOR ISOLATED REPRESENTATION OF SYNTACTIC AND SEMANTIC KNOWLEDGE

Holger Stahl, Johannes Müller

Institute for Human-Machine-Communication
Munich University of Technology
Arcisstrasse 21, D-80290 Munich, Germany
email: {sta,mue}@mmk.e-technik.tu-muenchen.de

ABSTRACT

A new form of a grammar is described, which provides two separate sets of stochastic parameters for representing both the semantic and the syntactic knowledge, required for automatic speech understanding. The semantic structure is introduced as an adequate representation of natural spoken, one-sentence command utterances. The constraints and probabilities delivered by the grammar can be integrated into the framework of a stochastic top-down parser to decode the semantic content of an utterance directly from its observation sequence. The performance of the developed methods is proved for the domain of a speech understanding graphic editor, which can be controlled solely by natural spoken commands.

Keywords: speech understanding, context-free grammar, stochastic models, syntactic and semantic knowledge

1. INTRODUCTION

1.1 A Speech Controlled Application

Using natural speech as a tool to operate a technical system [8], the semantic content of the utterance has to be found, which means speech *understanding*.

We chose a graphic editor (fig. 1) as a suitable application for understanding command utterances. The user should be facilitated to create, alter or delete three-dimensional objects like spheres, cuboids, cones or cylinders. We demand, that the speech understanding part of the system is able to analyse utterances, which are

- natural spoken sentences without subordinate clauses
- in continuous speech from an unknown speaker.

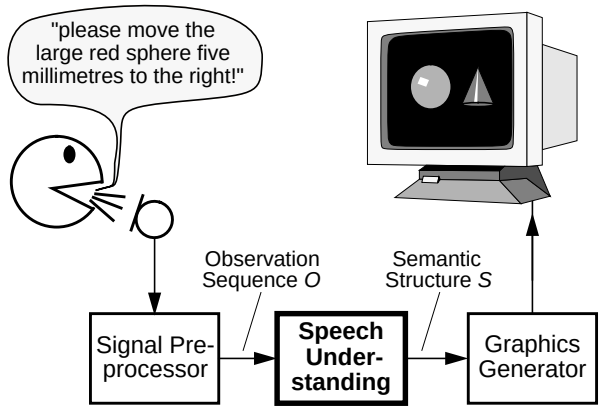

Figure 1: Project: A speech understanding graphic editor

1.2 Maximum-a-Posteriori Top-Down-Decoding

Stochastic methods have proved to be a powerful approach for speech *recognition* [6] (i.e. finding the most likely word chain given the utterance), so it is obvious to solve speech *understanding* in a similar way, too. Thus, the problem of mapping a sequence of observation vectors O to its semantic content S can be expressed by maximizing the maximum-a-posteriori probability $P(S|O)$:

$$S_E = \underset{S}{\operatorname{argmax}}\, P(S|O)\,. \tag{1}$$

Applying *Bayes*' inversion formula and taking into account just the most likely word chain W, we obtain the following classification rule, derived more detailed in [12]:

$$S_E = \underset{S}{\operatorname{argmax}} \max_{W} \left[P(O|W)\cdot P(W|S)\cdot P(S)\right]\,. \tag{2}$$

The probabilities $P(S)$ and $P(W|S)$ in eq. (2) have to be delivered by the grammar, which is described in the following chapters. The emission probability $P(O|W)$ is calculated by phoneme-based, continuous Hidden-Markov-Models, trained speaker independently, which can be adopted from existing speech recognition systems [5].

In contrast to the bottom-up strategy, which is applied for the speech understanding systems of many research groups ([2], [3], [13], and others) fig. 2 shows a top-down arrangement of the decoding process satisfying eq. (2). The top-down approach is well-established in speech recognition, and our own tests with a 'time of day'-understanding task gave encouraging results [1].

2. DEMANDS OF THE GRAMMAR *G*

The grammar G providing the a-priori probability $P(S)$ and the conditional probability $P(W|S)$ has to meet the following requirements:

- G contains rules, which are interpreted as stochastic events. These rules can be separated into two sets of rules contributing either to the a-priori probability $P(S)$ or to the conditional probability $P(W|S)$. In the following, this two sets of rules are called the *semantic model* G_{sem} and the *syntactic model* G_{syn}, respectively.
- A sequence of rule applications to originate a word chain is called a derivation. The most likely derivation to compose a certain word chain W is demanded to be equivalent to the semantic content S of the utterance. Only that rules in the derivation are significant, which are contained in the semantic model G_{sem}.

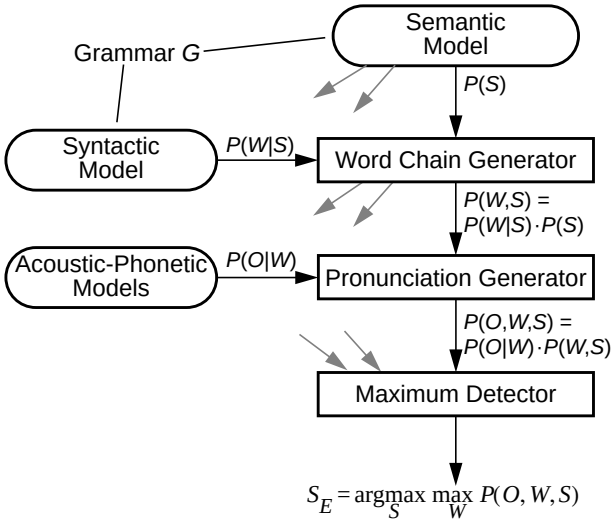

Figure 2: Layers of the speech understanding system

3. THE SEMANTIC MODEL $G_{\mathbf{sem}}$

3.1 Definition of the Semantic Structure

In [11], the semantic content of an utterance is assembled from "conceptual labels", which each express a small semantic partition of the utterance. In our approach, the *semantic structure S* (representing the semantic content) is a tree consisting of a finite number N of semantic units (we simply call them *semuns*) s_n:

$$S = \{s_1, s_2, \ldots, s_N\} \tag{3}$$

Each semun $s_n \in S$ with $1 \le n \le N$ is an (X+2)-tupel of a type $t[s_n]$, a value $v[s_n]$ and X *successor*-semuns[1] $q_1[s_n], \ldots, q_X[s_n] \in \{s_2, \ldots, s_N, \text{blnk}\} \setminus \{s_n\}$:

$$s_n = \Big(t[s_n], v[s_n], q_1[s_n], \ldots, q_X[s_n]\Big),\ X \ge 1 \tag{4}$$

The semun s_1 is defined as the root of the semantic structure S. Every semun $s_2, \ldots, s_N$ is marked exactly once as a successor semun. The special semun 'blnk' has the type $t[\text{blnk}] = \text{blnk}$, no value and no successor.

In the sense of predicate logic, a semun with X successors can be compared to an X-place relational constant [4]. In this context, a 0-place relational constant can be realized by a semun s_n with X=1 successor $q_1[s_n]$ and the successor type $t[q_1[s_n]] = \text{blnk}$.

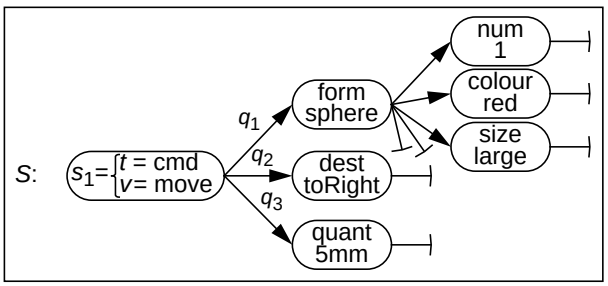

Figure 3: Semantic structure S in a graphic form

[1] Currently, we are using semuns with $1 \le X \le 5$ successors.

As an example, fig. 3 shows the semantic structure S of the German word chain "*bitte schiebe die große rote kugel fünf millimeter nach rechts*" ("*please move the large red sphere five millimetres to the right*").

3.2 Probabilities in the Semantic Model

If statistical dependencies are assumed only inside of each semun, $P(S)$ can be calculated as product of the following first order probabilities:

$$P(S) = f_{\text{root}} \cdot \prod_{n=1}^{N} (e_n \cdot f_n) \ , \text{with...} \tag{5}$$

- ... f_{root} denoting the a-priori probability that the root semun s_1 is of the type $t[s_1]$:

$$f_{\text{root}} = P(t[s_1]) \tag{6}$$

  The semantic model has to provide f_{root} for all types.

- ... e_n denoting the conditional probability that the value $v[s_n]$ occurs with the semun s_n of type $t[s_n]$:

$$e_n = P\Big(v[s_n] \,\Big|\, t[s_n]\Big) \tag{7}$$

  The semantic model has to provide this probability for all combinations of types $t[s_n]$ and values $v[s_n]$.

- ... f_n denoting the conditional probability that the X successor semuns $q_1[s_n], \ldots, q_X[s_n]$ of the semun s_n with type $t[s_n]$ are of the types $t[q_1[s_n]], \ldots, t[q_X[s_n]]$:

$$f_n = P\Big(t\big[q_1[s_n]\big], \ldots, t\big[q_X[s_n]\big] \,\Big|\, t[s_n]\Big) \tag{8}$$

  The semantic model has to provide this probability for all combinations of types $t[s_n]$ and the types of possible successors $t[q_x[s_n]]$, $x = 1, \ldots, X$.

4. THE SYNTACTIC MODEL $G_{\mathbf{syn}}$

4.1 Simplifying Assumptions

We assume the following restrictions for the word chains $W = w_1 w_2 \ldots w_j \ldots w_J$ originated by the syntactic model to express a given semantic content S:

- Every word w_j in the word chain W can be assigned to exactly one semun $s_n \in S = \{s_1, \ldots, s_N\}$.
- For each semun $s_n \in S$, one word w_{sig} is produced obligatorily, which depends on the value $v[s_n]$ of the semun s_n. Another word w_{insig} is produced optionally, which depends only on the type $t[s_n]$. We call these two words the *significant word* and the *insignificant word*, respectively.
- An unbroken part $w_i w_{i+1} \ldots w_j$ of W is originated for each semun $s_n \in S$ and all its successor branches.

4.2 Production Rules

The last assumption above implies to use a stochastic context-free grammar [10] as syntactic model. This grammar, denoted $G_{\text{syn}} = (V, T, \Sigma, P)$ contains the sets V, T and P of variables, terminals and production rules. The

derivation always starts rewriting the start symbol $\Sigma \in V$ as variable $A(s_1)$, with s_1 marking the root of the semantic structure S:

$$P\Big(\Sigma \rightarrow A(s_1)\Big) \tag{9}$$

The probabilities for rewriting the variables $A(s_n)$, $B(s_n)$ and $C(s_n)$ each depend on the characteristics of one explicit semun $s_n \in S$:

- For the case $s_n \neq \text{blnk}$, the variable $A(s_n)$ produces a sequence of a variable $B(s_n)$, an optional variable $C(s_n)$ and X variables $A(q_1[s_n]), \ldots, A(q_X[s_n])$ for the successors of s_n:

$$P\Big(A(s_n) \rightarrow \underbrace{B(s_n), \overbrace{C(s_n)}^{\text{optional}}, A(q_1[s_n]), \ldots, A(q_X[s_n])}_{\text{in arbitrary order}} \Big| t[s_n]\Big) \tag{10}$$

  $A(\text{blnk})$ always produces the empty string ε:

$$P\Big(A(\text{blnk}) \rightarrow \varepsilon\Big) = 1 \tag{11}$$

  The probability of eq. (10) has to be provided by the syntactic model for all imaginable arrangements of the above values depending on all possible types $t[s_n]$ of the semun s_n. It is estimated by a transition network similar to an ergodic hidden markov model (fig. 4). Such a *syntactic module* (SM) consists of $X+4$ states: $B(s_n)$, $C(s_n)$ and $A(q_1[s_n]), \ldots, A(q_X[s_n])$ represent the corresponding variables, 'strt' and 'end' stand for the entry and the exit of the SM. The probability for the arrangement (eq. (10)) is approximated by multiplying all transition probabilities $a_{x,y}$ along a certain path through the SM. This path, i.e. the order of passing the states of the SM is constrained by eq. (10). The syntactic model has to provide a separate set of SM-parameters $a_{x,y}$ for all types $t[s_n]$ of the semun s_n.

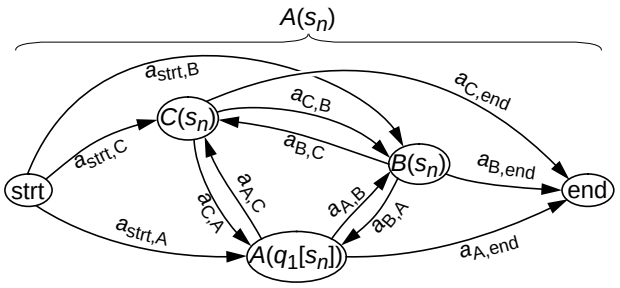

Figure 4: SM for the semun s_n with $X = 1$ successor

- $B(s_n)$ produces one significant word w_{sig} out of the vocabulary depending on the value $v[s_n]$ of the semun s_n:

$$P\Big(B(s_n) \rightarrow w_{\text{sig}} \Big| v[s_n]\Big) \tag{12}$$

- $C(s_n)$ produces one insignificant word w_{insig} out of the vocabulary depending on the type $t[s_n]$ of s_n:

$$P\Big(C(s_n) \rightarrow w_{\text{insig}} \Big| t[s_n]\Big) \tag{13}$$

The desired probability $P(W|S)$ is calculated by maximizing the product of the probabilities concerning all the productions according to eq. (10), (12) and (13) required to derive '$\Sigma \Rightarrow W$'.

5. INTEGRATION OF THE SEMANTIC AND THE SYNTACTIC MODEL

The stochastic process of originating word chains with the grammar described above can be seen as a complex transition network. The syntactic model G_{syn} is represented by a set of syntactic modules (SMs) according to fig. 4., the semantic model G_{sem} connects these SMs with transition edges. Fig. 5 depicts one selected path through such a network, consisting of four SMs:

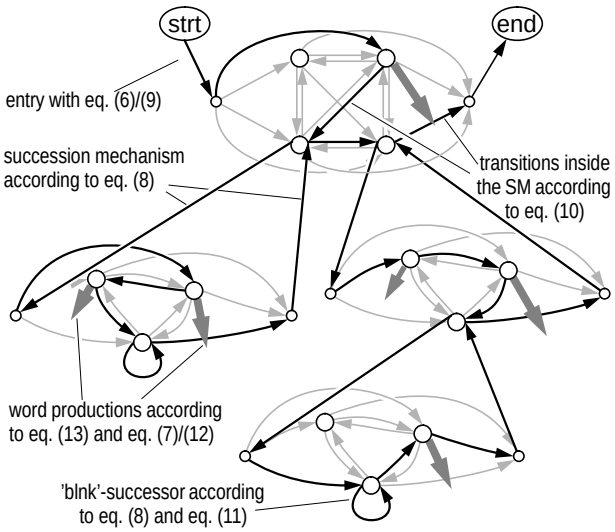

Figure 5: Origination of a word chain along a search path as a sequence of transitions within the grammar network initiated by the semantic and the syntactic model

- The first syntactic module representing the semun s_1 of the semantic structure S is entered with the probability f_{root} in eq. (6). So the type $t[s_1]$ is fixed.
- For every further semun s_n inserted into the semantic structure, a new SM is entered with the probability f_n according to eq. (8). So the semun's type $t[s_n]$ is fixed.
- The time alignment of the emitted words is taken out by passing through each SM with the probability and the order according to eq. (10).
- From the state $C(s_n)$ an insignificant word w_{insig} is emitted with the probability according to eq. (13).
- The value $v[s_n]$ of each semun s_n effects the significant word w_{sig}, which is emitted from state $B(s_n)$. Hence, before the word is emitted according to eq. (12), the value has to be fixed with the probability e_n according to eq. (7).

6. EVALUATION AND RESULTS

As training and testing data, 1843 utterances to operate a simple graphic editor (fig. 1) were collected in a Wizard-of-Oz simulation from 33 different speakers [9]. The utterances contain about 7 words on average. Training was taken out by a strategy similar to the Inside-Outside algorithm shown in [7], based on counting the number of executed rules in an iterative process. The number of types and values in the models was 41 and 250, respectively, the resulting number of model parameters is about 2500.

To evaluate the grammar, the utterances' semantic contents were extracted according to eq. (2). However, the acoustic-phonetic modelling problem was left aside at this time, so the word chain W itself was the input of the parser and the factor $P(O|W)$ in eq. (2) was omitted. Fig. 2 outlines the cooperation of the training and evaluation processes and explains our definition of the 'performance rate':

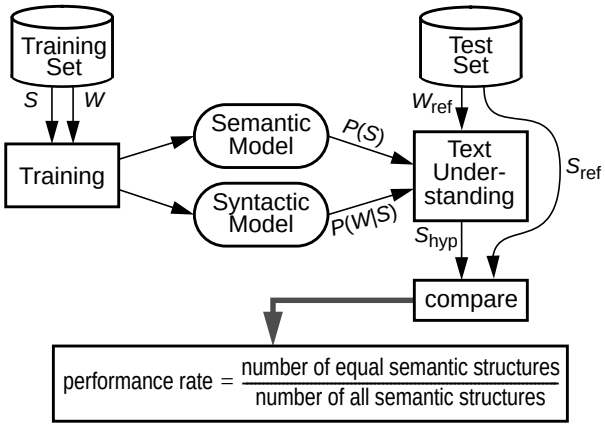

Figure 6: Evaluation of the performance rates

Tab. 1 shows the performance of the text-understanding system for equal training and testing material.

We distinguish models with

- continuous probabilities $0 \le P(\ldots) \le 1$ and
- discrete probabilities, which means that all probabilities in the respective model are set either 0 or 1.

The performance increases significantly by using models with continuous probabilities, which shows the importance of stochastic production rules in the grammar:

1843 equal training and test utterances		
semantic model	syntactic model	performance rate
continuous	continuous	99.8 %
continuous	discrete	98.8 %
discrete	continuous	96.1 %
discrete	discrete	93.0 %

Table 1: Percentage of correctly assigned semantic structures with identical training and test set

Tab. 2 shows the performance of the text understanding system for disjoint training and testing material. In this case, 16,3% of the 92 utterances have been rejected due to words, which have not been seen in the training, 5.4% have been rejected due to their syntax. Only 1.1% (1 utterance) was really misrecognized!

1751 training utterances and 92 test utterances		
semantic model	syntactic model	performance rate
continuous	continuous	77.2 %

Table 2: Percentage of correctly assigned semantic structures with disjoint training and test sets

7. CONCLUSIONS

The presented grammar is suitable for accurate understanding text of simple formed command utterances. The problem of lacking vocabulary may be reduced by automatically adding a number of word equivalents from a dictionary of synonyms.

The integration of the acoustic-phonetic models delivering $P(O|W)$ into the framework of the top-down approach has been already completed. Presently, we are optimizing performance rates and computation effort, simultaneously we develop the graphics generator (fig. 1) for converting semantic structures into a graphic database access. The search algorithm as well as the training algorithms will be presented in the future.

REFERENCES

[1] J.G. Bauer, H. Stahl, J. Müller: *A One-Pass Search Algorithm for Understanding Natural Spoken Time Utterances by Stochastic Models*, Proc. EUROSPEECH 1995 (Madrid, Spain), to be published

[2] M.K. Brown, B.M. Buntschuh: *A Context-Free Grammar Compiler For Speech Understanding Systems*, Proc. ICLSP 1994, (Yokohama, Japan), pp. 21-24

[3] W. Eckert et al.: *A Spoken System for German Intercity Train Timetable Inquiries*, Proc. EUROSPEECH 1993 (Berlin, Germany), pp. 1871-1871

[4] G. Görz (ed.): *Einführung in die künstliche Intelligenz*, Addison-Wesley, 1993

[5] A. Hauenstein, E. Marschall: *Methods for Improved Speech Recognition over Telephone Lines*, Proc. IEEE ICASSP 1995 (Detroit, Michigan USA), pp. 425-428

[6] H. Höge: *Statistische Modelle für die Spracherkennung*, Proc. DAGA 1993 (Frankfurt am Main, Germany), pp. 11-30

[7] F. Jelinek, J.D. Lafferty, R.L. Mercer: *Basic Methods of Probabilistic Context Free Grammars*, Proc. NATO ASI, vol. F75, Springer, 1992, pp. 345-360

[8] M. Lang, H. Stahl: *Spracherkennung für einen ergonomischen Mensch-Maschine-Dialog*, mikroelektronik, vol. 8 (1994), no. 2, pp. 79-82

[9] J. Müller, H. Stahl: *Collecting and Analyzing Spoken Utterances for a Speech Controlled Application*, Proc. EUROSPEECH 1995 (Madrid, Spain), to be published

[10] H. Ney: *Stochastic Grammars and Pattern Recognition*, Proc. NATO ASI, vol. F75, Springer, 1992, pp. 319-344

[11] R. Pieraccini, E. Levin, E. Vidal: *Learning how to Understand Language*, Proc. EUROSPEECH 1993 (Berlin, Germany), pp. 1407-1412

[12] H. Stahl, J. Müller: *An Approach to Natural Speech Understanding Based on Stochastic Models in a Hierarchical Structure*, Proc. Workshop Modern Modes of Man-Machine-Communication (Maribor, Slovenia) 1994, pp. 16/1-16/9

[13] M. Woszczyna et al.: *Recent Advances in Janus: A Speech Translation System*, Proc. EUROSPEECH 1993 (Berlin, Germany), pp. 1295-1298