BELIEF NETWORKS FOR A SYNTACTIC AND SEMANTIC ANALYSIS OF SPOKEN UTTERANCES FOR SPEECH UNDERSTANDING

Marc Hofmann and Manfred Lang
Institute for Human-Machine Communication
Technical University of Munich, D-80290 Munich, Germany
{hofmann, lang}@ei.tum.de

ABSTRACT

In this paper we present a new approach towards speech understanding that merges semantic and intention decoding to one component. The algorithm is supposed to evaluate a speech recognizer's utterance hypotheses regarding a) syntactical and semantical relations between words and phrases and b) potential intentions of the user. The mathematical fundament for this evaluation is probability theory. We make use of belief networks to handle the analysis of an utterance hypothesis as a process of reasoning with uncertain and incomplete information. The algorithm in general can be characterized as phrase spotting.

The algorithm proved to be very robust for controlling the navigation system and the audio equipment of a car.

1. INTRODUCTION

Speech controlled applications are often based on word spotting for interpreting the user's utterances. Obviously word spotting is not very robust at noisy signals or if the keyword is part of several intentions, but on the other hand word spotting is able to deal with the out-of-vocabulary problem and with grammatically incorrect utterances. Classical approaches towards speech understanding use syntactic and semantic relations between words and phrases making recognition of the user's intention more robust, but also susceptible for out-of-vocabulary phenomena and grammatically incorrect utterances. In this paper we introduce an approach which combines the advantages of word spotting with the advantages of classical speech understanding. It can be interpreted as phrase spotting. Therefore we make an extensive use of belief networks for a syntactical and semantical evaluation of utterance hypotheses. The paper is structured as follows. First we give a short introduction to belief networks. We will then explain the algorithm, first in principle, then in detail. The paper will end with results and conclusions.

2. BELIEF NETWORKS

Methods based on probability theory seem to be the state-of-the-art technique for speech recognition on the signal level, for example Hidden Markov Models. We use another method of probabilistic reasoning for interpreting an utterance, namely belief networks. In this paper we can only provide a very brief description of belief networks.

A belief network consists of a set of nodes, which are related to state variables $\boldsymbol{X}$, each with a finite set of states. Directed edges between the nodes express statistical dependencies between the related state variables, quantitatively expressed as conditional probabilities of nodes and their parent nodes. The joint probability distribution provides a complete representation of network structure and conditional probabilities. The joint probability distribution of a belief network consisting of n random variables can be easily calculated as follows [1][2]:

$$P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid parents\ (X_i)) \qquad (1)$$

A belief network provides methods of inferring the states of some query variables based on observations regarding the evidence variables.

With their ability to deal with uncertain and incomplete information, belief networks are the mathematical background of the approach we present in the next section.

3. METHODS

3.1 Intention-Based Evaluation

The application we want to control by spontaneous speech is the IT-equipment of a car, namely the navigation system and the audio equipment. To show the problems arising if a user talks quite fast, maybe omitting some word endings and maybe using some out-of-vocabulary words, Fig. 1 shows a speech recognizer's output of 130 utterance hypotheses ordered according to their confidence measures. The utterance has been spoken quite fast, without speaking explicitly understandable and with the German word "doch" being not part of the speech recognizer's vocabulary. The example is based on German language because our system has a German vocabulary and a literal translation of the speech recognizer's results into English isn't reasonable as the utterance hypotheses are semantically incorrect. Comparing the pronounced utterance with the most likely utterance hypothesis, the difference is quite considerable and will obviously cause problems for intention decoding. The 130th hypothesis differs even more. Looking at all 130 utterance

From: Proc. ICSLP 2000, IEEE

hypotheses, the 22nd is the one which is closest to the original, despite not being afflicted with the maximum confidence measure. This phenomenon is to be ascribed to the wrong segmentation of some words because of omitted endings. Our algorithm combines speech understanding and intention decoding, it is supposed to replace the speech recognizer's confidence measure for whole utterances by a new evaluation measure, based on syntactic and semantic considerations.

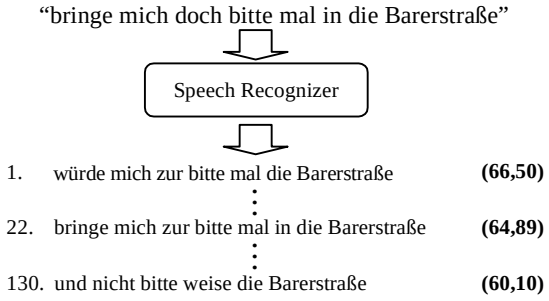

Figure 1: This figure shows the result of a speech recognition process with a German utterance as input. On the right the confidence measure of each utterance hypothesis is shown.

Fig. 2 shows the principle of our algorithm. The algorithm requires knowledge about every potential intention and the words and phrases to pronounce it. These informations are stored in the intention library which contains all *m* intentions to reason about. All *n* utterance hypotheses will be compared to each intention resulting in a quantitative evaluation measure for each hypothesis. Fig. 2 refers to the evaluation measure of the y^{th} utterance hypothesis regarding the x^{th} intention as EM_{xy}. The utterances with the maximal evaluation measure is the one that fits best to the related intention. This has to be done for all *m* intentions. The most likely intention is the one with the utterance afflicted with the overall maximal evaluation measure. The key of the algorithm is the syntactic-semantic evaluation component which will be described in the following sections.

3.2 Phrase Representation

All intentions of the intention library are related to utterances which could be used for controlling the navigation and the audio system of a car. For such quite simple applications an utterance normally has two purposes: to tell the system which system parameter to change and to tell the system the new parameter. Therefore we divide an utterance into two types of phrases:

- *operator phrase*: to tell the system which system parameter to change
- *parameter phrase*: to tell the system the new value of the parameter

The first step is to divide all utterances of an intention into operator and parameter phrases. Fig. 3 gives an example of splitting utterances into operator and parameter phrases.

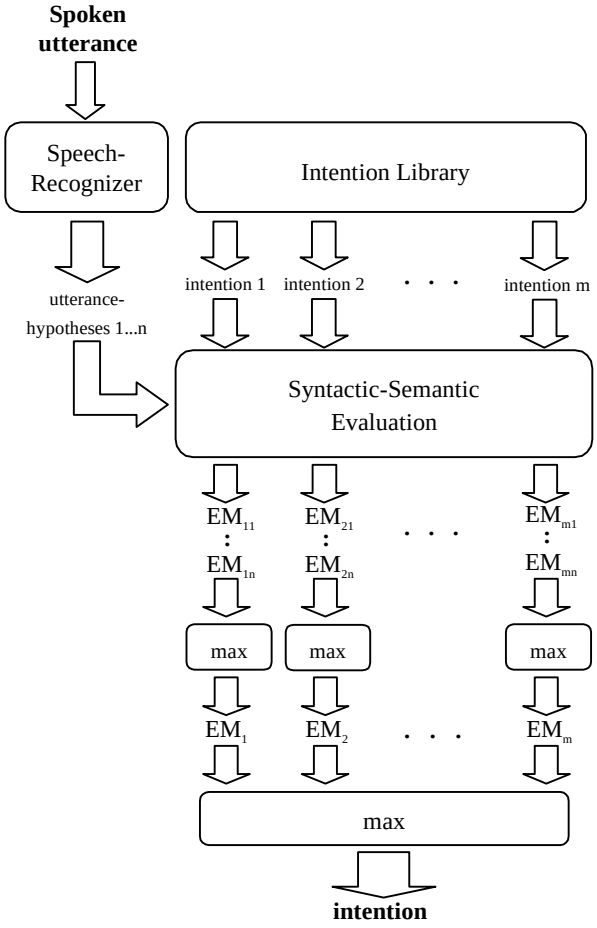

Figure 2: This graph shows the main components of the algorithm.

Grammatically and semantically irrelevant words will not be considered. The classification has to be done in such a way, that every possible combination of operator and parameter phrase is syntactically and semantically correct.

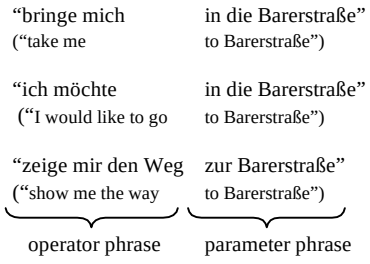

Figure 3: Example for splitting utterances into operator and parameter phrases.

The fact that an utterances consists of an operator phrase and a parameter phrase is being modeled by a belief network (Fig. 4). The three nodes are related to boolean state variables. The phrase variables represent the observation of the respective phrase. The conditional probabilities are chosen according to a

logical AND-function, i.e. *Intention = Operator ∧ Parameter*. An utterance is only completely observed if an operator and a related parameter phrase is completely observed.

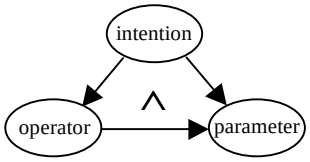

Figure 4: The structure of an intention network consisting of operator and parameter node.

This kind of belief network has to be used for each intention to reason about with the probability of the *intention* node being the evaluation measure. We will refer to it as *intention network*.

For every phrase a subnetwork with one boolean node for each word, structured and trained analogically to the intention networks, will be created. The probability of the root node is supposed to give an indication of how well the observed words fit to that phrase.

Fig. 5 shows a representation of an intention consisting of operator and parameter phrase networks. Combining operator and parameter phrases creates utterances for expressing the respective intention.

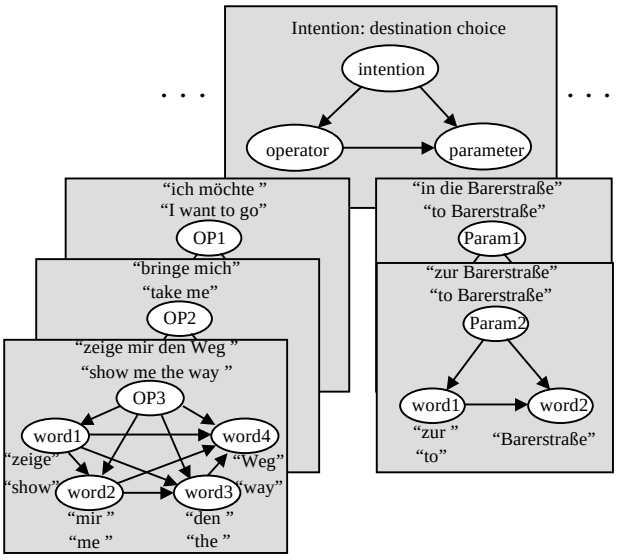

Figure 5: Every intention is represented by an intention network and a set of operator and phrase networks. The words related to the word nodes are pictured below the nodes in German and in English.

After modeling every intention of the intention library with an intention network and a set of phrase networks, the evaluation process can start.

3.3 Syntactic-Semantic Evaluation

For better illustration of the syntactic-semantic evaluation of an utterance hypothesis Fig. 6 shows one intention network with two corresponding phrase networks; the algorithm of course includes all intentions with their subnetworks. At the bottom of Fig. 6 the utterance hypothesis to evaluate is pictured. The algorithm will be explained in the following 8 steps (see Fig. 6 ① to ⑧) :

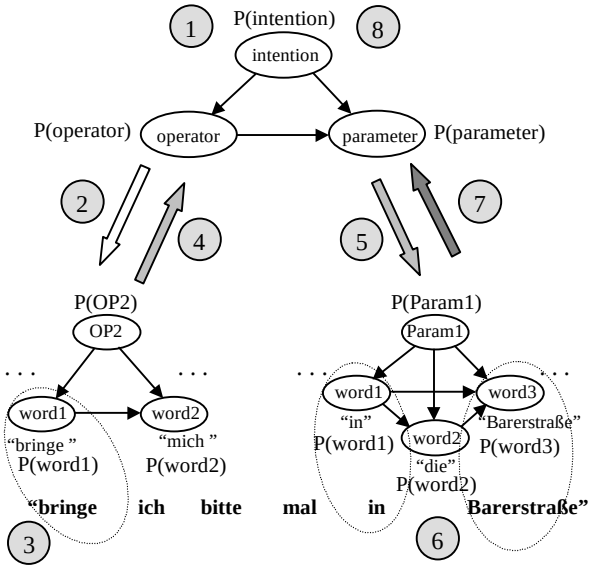

Figure 6: Illustration of the syntactic-semantic evaluation of an utterance hypothesis regarding a certain intention.

① At the beginning the *intention* node of the intention network is assigned a neutral a-priori probability distribution, i.e. both states are equally likely. Given an a-priori probability for the root node the marginal probabilities for the phrase nodes can be calculated.

② To coordinate the operator networks with the intention network their root nodes are assigned the probability of the *operator* node *P*(*operator*). That exerts influence on the word nodes, changing their marginal probabilities.

③ The utterance hypothesis is parsed for words related to the word nodes. In Fig. 6 the word “bringe” is part of the utterance hypothesis and part of the vocabulary of the node *word1*. This observation has to be mapped on the network. The evaluation of an utterance hypothesis is based on its single words with their confidence measures assigned by the speech recognizer. We will treat such confidence measures as uncertain information. Therefore the standard inference algorithm for belief networks has to be modified.

Basically inference algorithms for belief networks allow reasoning about query variables given the states of the evidence variables. This is quite an easy task if the joint probability is known, as we just have to search for the entries

of the joint probability table according to the states of the evidence variables.

Now we have the situation that we are just able to make uncertain, probability based statements about the evidence variables. Therefore the inference algorithm has to be extended. According to the Bayesian theorem the joint probability of the operator phrase network can be expressed as follows:

$$P(OP\,2, word\,1, word\,2) = P(OP\,2, word\,2 \mid word\,1) \cdot P(word\,1) \quad (2)$$

Changing the belief of observing a word, that means making a probability based statement, will also influence the joint probability resulting in a modified joint probability P_{new}:

$$P_{new}(OP\,2, word\,1, word\,2) = P(OP\,2, word\,2 \mid word\,1) \cdot P_{new}(word\,1) \quad (3)$$

Dividing the equations (2) and (3) results in equation (4):

$$P_{new}(OP\,2, word\,1, word\,2) = P(OP\,2, word\,1, word\,2) \cdot \frac{P_{new}(word\,1)}{P(word\,1)} \quad (4)$$

This equation enables the inference algorithm to deal with changes in belief. It will be used to treat the words of an utterance hypothesis as a change in belief of observing a special word.

Generally the marginal probability of a word node of a phrase network reflects the assumption that one word of the related vocabulary is part of the utterance hypothesis. This has to be seen from a syntactic and semantic point of view as a change of the probability of a word node will also affect the other word nodes of the phrase. The next step is to merge the marginal probability of the node *word1* with the confidence measure of the speech recognizer. We interpret the range from *P*(*word1*) to 1 as space which is left for observations by the speech recognizer. This means the confidence measure will be mapped on that range as contribution by the speech recognizer and added to *P*(*word1*), which express the impact of syntax and semantics within a phrase and previous observations. As the range of the confidence measure is from 0 to 100 we have to normalize it. Equ. (5) gives a mathematical description of the above:

$$P_{ob}(word\,1 = y) = P(word\,1 = y) + \frac{c}{100}(1 - P(word\,1 = y)) = = P(word\,1 = y) + \frac{c}{100} P(word\,1 = n) \quad (5)$$

The resulting new marginal probability P_{ob}, which stays abreast of syntactic-semantic considerations and of the uncertainty in the output of the speech recognizer, will be entered into the network as uncertain information (Equ. (4)). The more words of a phrase have been observed, the higher is the probability of observing other words of that phrase, that means the expectation increases. Hence syntactically and semantically related words will be emphasized.

The new probability P_{ob}(*word1*) is entered into the network according to Equ. (4). Making use of belief networks' ability to deal with incomplete information, word nodes with unobserved words remain not instantiated. The above procedure has to be done for all operator phrases of the current intention.

④ The operator phrase *x* with the highest root node probability *P*(*OPx*) has to be determined. This is the phrase that is described most completely by the hypothesis utterance regarding the current intention. *Max* {*P*(*OPx*)} will now be entered into the intention network by assigning it to the *operator* node of the intention network (Equ. (4)). This will also affect the *intention* and the *parameter* node, emphasizing syntactically and semantically related parameter phrases.

⑤ By analogy to step ② the marginal probability *P*(*parameter*) will be assigned to the root node of every parameter network.

⑥ By analogy to step ③ the hypothesis utterance is parsed for words of the vocabularies of the phrase's word nodes. Common words have to be mapped on its word nodes.

⑦ By analogy to step ④ the parameter phrase *y* with the highest probability has to be determined *P*(*Paramy*). *Max* {*P*(*Paramy*)} will be entered into the intention network.

⑧ The probability of the *intention* node of the intention network is the quantitative measure of how well an utterance hypothesis fits an intention. Therefore *P*(*intention*) is the evaluation measure EM for the utterance hypothesis concerning the intention to reason about.

4. RESULTS & CONCLUSIONS

The presented algorithm has been implemented and evaluated. As mentioned before, the application to control is a navigation system and the audio equipment of a car. The algorithm has proved to be very successful for this kind of application with a recognition rate of the user's intention of about 90 per cent. It is able to cope with spontaneous, natural speech as well as with command language. To some extent it shows robustness regarding out-of-vocabulary words.

The current algorithm performs very well in applications with simply structured utterances. More complex utterances, for example with relative clauses, will entail problems for mapping observations on the phrase networks, as the current algorithm is not able to deal with several observations for one word node. A solution of that problem is left for future work and will generally enhance the algorithm's performance in simple and in more complex domains.

5. REFERENCES

1. Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988

2. Russell, S., and Norvig, P., *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.