TRAINING OF THE SEMANTIC MODEL FOR A SPEECH UNDERSTANDING GRAPHIC EDITOR

Janez Kaiser [a)], Johannes Müller [b)], Holger Stahl [b)]

a) Faculty of Electrical Engineering and Computer Science, Institute of Electronics
Smetanova 6, 62000 Maribor, Slovenia

b) Institute for Human-Machine-Communication, Munich University of Technology
Arcisstrasse 21, D-80290 Munich, Germany

Abstract

In this paper, an algorithm for the training of the semantic model and the addition of floor values into the semantic model for a speech controlled graphic editor is presented. The speech understanding module of the graphic editor uses a grammar, consisting of two stochastic knowledge bases: the semantic model and the syntactic model. Within a 'top-down' strategy, the semantic model generates semantic structures, which are semantic representations close to the word level, with the respective probabilities.

1 Introduction

Speech understanding can be interpreted as mapping of an observation sequence O to its semantic content, represented by the semantic structure S. Given the observation sequence O, the most likely S_E has to be found. Due to the high variety of O and S, additional representation levels are necessary. Clearly defined is the word level W. The problem of finding the most likely semantic content S_E can be written as follows [4]:

$$S_E = \underset{S}{\operatorname{argmax}} \max_{W} [P(O|W) \cdot P(W|S) \cdot P(S)] \qquad (1)$$

- The semantic model delivers the a-priori probability $P(S)$ for the occurrence of a semantic structure S.
- The syntactic model delivers the conditional probability $P(W|S)$ for the occurrence of a word chain W given a certain semantic structure S.
- The acoustic-phonetic model delivers the conditional probability $P(O|W)$ for the occurrence of an observation sequence O given a certain word chain W.

In this paper, we only consider the semantic model. The syntactic model is described in [6], the acoustic-phonetic model can be taken from existing speech recognition systems (e.g. SPICOS [1] or SPRING [7]).

The whole system is part of a speech understanding graphic editor (see figure 1), developed at the Institute for Human-Machine-Communication, Munich University of Technology.

2 Definition of the semantic structure

The semantic structure represents the semantic content of an utterance. Since this content can be very complex, the semantic structure S is divided into N smaller semantic units s_n, abbreviated semuns, which have limited variety [2]. Every semun corresponds to one significant word in the word chain.

$$S = \{s_1, s_2, \ldots, s_N\} \qquad (2)$$

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 $q_1[s_n], \ldots, q_X[s_n] \in \{s_2, \ldots, s_N, \text{blnk}\} \setminus \{s_n\}$:

$$s_n = (t[s_n], v[s_n], q_1[s_n], \ldots, q_X[s_n]), X \ge 1 \qquad (3)$$

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.

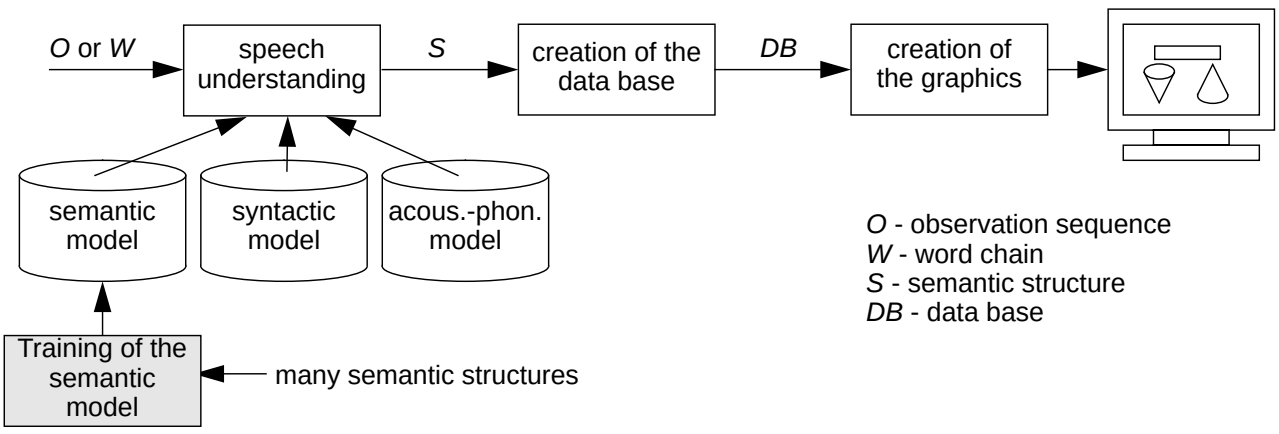

Figure 1: Structure of the graphic editor

- The type $t[s_n]$ of the semun s_n lays down the number X of possible types of successor semuns and makes selection of possible values of the semun s_n.
- The value $v[s_n]$ of the semun s_n shows the proper meaning of a significant word, which corresponds to the semun s_n

As an example, figure 2 shows the word chain *W* and the semantic structure *S* of an utterance.

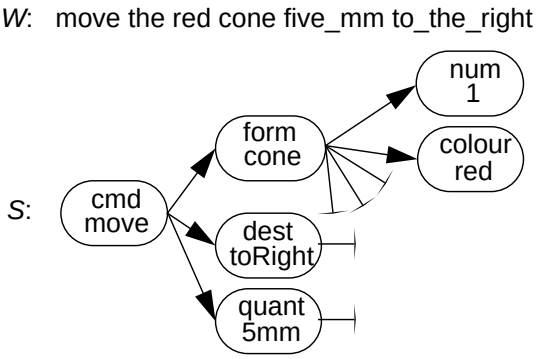

Figure 2: Word chain and semantic structure

3 Probabilities in the semantic model

The variety of useful semantic structures is too large to estimate their probability $P(S)$ directly from a training corpus. Therefore, some first order probabilities are defined:

- The **root probability** f_{root} is the a-priory probability that the root semun s_1 is of the type $t[s_1]$:

$$f_{\text{root}} = P(t[s_1]) \tag{4}$$

  The semantic model has to provide the probability f_{root} for all types of semuns.
- The **value probability** e_n is the a-priory probability that the semun s_n of the type $t[s_n]$ has the value $v[s_n]$:

$$e_n = P\left(v[s_n] \,\middle|\, t[s_n]\right) \tag{5}$$

The semantic model has to provide the probability e_n for all combinations of types and values.

- The succession probability f_{n} is the conditional probability that 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 $t[q_1[s_n]], \ldots, t[q_X[s_n]]$:

$$f_n = P\left(t[q_1[s_n]], \ldots, t[q_X[s_n]] \mid t[s_n]\right) \tag{6}$$

  The semantic model has to provide the probability f_{n} for all combinations of types and possible successors.

If statistical independence of these terms is assumed, $P(S)$ can be calculated as follows [5]:

$$P(S) = f_{\text{root}} \cdot \prod_{n=1}^{N} (e_n \cdot f_n) \tag{7}$$

4 Training of the semantic model

In order to estimate the probabilities in the semantic model, two procedures are applied on the training material, the initialization and the iteration, see figure 3.

Initialization

From the given training material, the defined root, value and succession probabilities are calculated by determining the frequency of occurrences within the whole training material.

Iteration

The goal of the iteration is to optimize the semantic model for the given training data. In each semantic structure, the

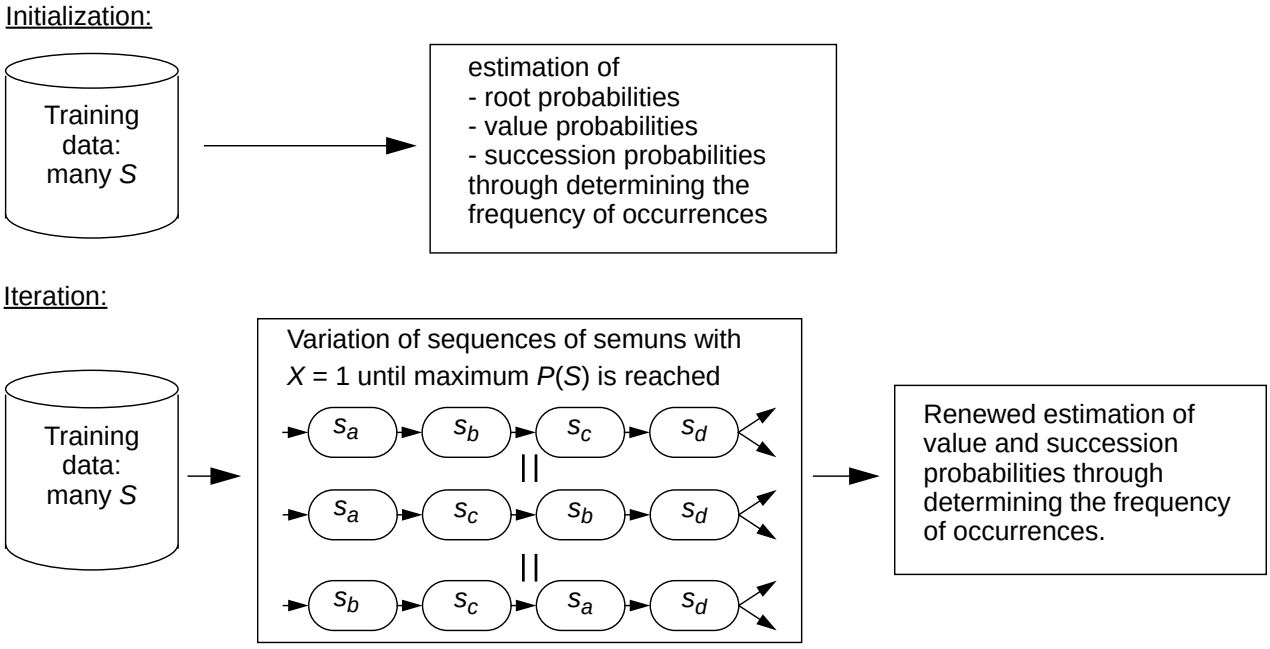

Figure 3: Training of the semantic model

sequences of semuns with exactly one successor are varied until the maximum $P(S)$ is reached. Even though the semantic structure is changed in this process, the semantic content of the semantic structure remains unchanged. In each iteration step, all semantic structures in the training data are optimized and the parameters for a new semantic model are calculated. For the calculation of $P(S)$, which is used in the optimization process, the semantic model from the previous iteration step is used (for the first iteration step, the initialised model is taken). The iteration is carried out as many times until the new semantic model and the model from the previous iteration step are identical. When optimizing a semantic structure, the following steps are necessary:

1. All sequences of semuns with exactly one successor within the semantic structure are found. All sequences are optimized by carrying out the steps 2 to 6 for each sequence.
2. The number l of semuns in the sequence is determined.
3. The number v of possible variations is calculated:

$$v = l! \tag{8}$$

4. Each variation of the sequence is checked for correctness. Only some of the possible variations are allowed. The reason for this lays in the fact that the corresponding word chain has to remain emitable by the given syntactic model [6]. Only the allowed variations are used in optimization process.
5. For each variation, $P(S)$ is calculated. By varying the sequence of semuns, the succession probability of the predecessor semun of the first semun in the sequence and the succession probabilities of all semuns in the sequence change. Therefore also $P(S)$ changes.
6. The variation with the highest $P(S)$ is used in the optimized semantic structure.
7. A new semantic model based on optimized semantic structures is calculated.

5 Floor values

In the training material, only some of possible combinations of successor semun types for a given semun can be found. To enable the recognition of the combinations, which are not incorporated in the training data, the floor values are introduced. The missing combinations are added to the semantic model and their succession probability is set to be the lowest succession probability found for the existing combinations.

Example:

In a semantic structure, let the semun s_n have X=3 successors $q_1[s_n], q_2[s_n], q_3[s_n]$. In the training material, the following combinations of successor semun types have been found:

$$\{t_1[q_1], t_3[q_2], t_4[q_3]\},$$
$$\{t_2[q_1], t_3[q_2], t_5[q_3]\},$$
$$\{t_1[q_1], t_3[q_2], t_6[q_3]\}.$$

We have two possible types for q_1 ($t_1[q_1]$ and $t_2[q_1]$), one possible type for q_2 ($t_3[q_2]$) and three possible types for q_3 ($t_4[q_3]$, $t_5[q_3]$ and $t_6[q_3]$). Hence, $2 \cdot 1 \cdot 3 = 6$ possible combinations of successor semun types have to exist:

$$\left.\begin{array}{l}\{t_1[q_1], t_3[q_2], t_4[q_3]\}\\ \{t_2[q_1], t_3[q_2], t_5[q_3]\}\\ \{t_1[q_1], t_3[q_2], t_6[q_3]\}\end{array}\right\} \text{already existing}$$

$$\left.\begin{array}{l}\{t_1[q_1], t_3[q_2], t_5[q_3]\}\\ \{t_2[q_1], t_3[q_2], t_4[q_3]\}\\ \{t_2[q_1], t_3[q_2], t_6[q_3]\}\end{array}\right\} \text{added}$$

6 Results

Performance rates are defined as the ratio between the correctly understood semantic structures and all semantic structures used for evaluation. For evaluating the performance rates, the semantic model which included semuns with up to three successors was used. Semantic structures based on this model usually have sequences of semuns with exactly one successor.
Training and testing data consisted of word chains in German language. For each word chain, a correct corresponding semantic structure was available.

Iteration of the semantic model

For evaluation of the performance rates, training data consisting of 1843 word chains in German language [3] and corresponding semantic structures were used. The testing data were the same 1843 word chains with manually optimized semantic structures. Four iterations of the semantic model were needed until the optimum was reached.

Table 1: Performance rates

	only initialised	after 2 iterations	after 4 iterations
semantic structures	92.78 %	95.28 %	95.39 %
individual semuns	99.81 %	99.89 %	99.89 %

Table 1 shows that 2 iterations of the semantic model have brought 2.50% increase in performance rates of semantic structures and 0.08% increase in performance rates of individual semuns. After 4 iterations, 2.61% increase in performance rates of semantic structures and 0.08% increase in performance rates of individual semuns was achieved.

Floor values

For evaluation of the performance rates, training data consisting of 1659 word chains and corresponding semantic structures were used. The testing data consisted of 183 word chains and corresponding semantic structures. The semantic structures in the testing data were

not a subset of the training data. First, the semantic model was iterated four times and then the floor values were introduced.

Table 2: Performance rates

	after 4 iterations	after 4 iterations with floor values
semantic structures	66.12 %	66.67 %
individual semuns	79.87 %	79.94 %

Table 2 shows that the introduction of floor values has brought 0.55% increase in performance rates of semantic structures and 0.07% increase in performance rates of individual semuns. In this case, 17.49% of the word chains have been rejected due to missing words, which have not been seen in the training.

7 Conclusions

In this paper, the training of the semantic model and the introduction of floor values into this model were presented. The iteration of the semantic model is used to optimize this model according to the given training data. Floor values are introduced into the semantic model for different training and testing data. It can be seen that both procedures have brought increase in performance rates. Since both procedures take relatively short time to perform and bring considerable increases in performance rates, they prove to be useful.

References

[1] H. Höge: *SPICOS II - a Speech Understanding Dialogue System*, Proc. ICSLP 1990 (Kobe, Japan), pp. 1313-1316

[2] J. Müller, H. Stahl: *Die semantische Gliederung als adäquate semantische Repräsentationsebene für einen sprachverstehenden 'Grafikeditor'*, Proc. GLDV-Jahrestagung 1995 (Regensburg, Germany), to be published

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

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

[5] 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', 1994 (Maribor, Slovenia), pp. 16/1-16/9

[6] H. Stahl, J. Müller: *A Stochastic Grammar for Isolated Representation of Syntactic and Semantic Knowledge*, Proc. Eurospeech 1995 (Madrid, Spain), to be published

[7] K. Wothke et al.: *The SPRING Speech Recognition System for German*, Proc. Eurospeech 1989 (Paris, France), vol. 2, pp. 9-12