AUTOMATIC SPEECH TRANSLATION BASED ON THE SEMANTIC STRUCTURE

Johannes Müller, Holger Stahl, Manfred Lang

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

ABSTRACT

This paper describes a system for the semantic-based translation of spoken or written limited-domain utterances. The semantic structure as output of a semantic decoder serves as the interlingua-level. A word chain generator combined with a linguistic post-processor produces the according word chain in the target language. Both the semantic decoder and the word chain generator work with pure stochastic and trainable knowledge bases. The grammatical features of certain words can be easily extracted by the help of both the word chain and the semantic structure.

Keywords: automatic speech translation, speech understanding, semantic decoding, language production, semantic structure, semantic model, syntactic model, inflectional model

1. SYSTEM OVERVIEW

The depicted translation system, suitable for limited-domain utterances (comparable to [3] and [14]) without any subordinate clause, consists of three main components:

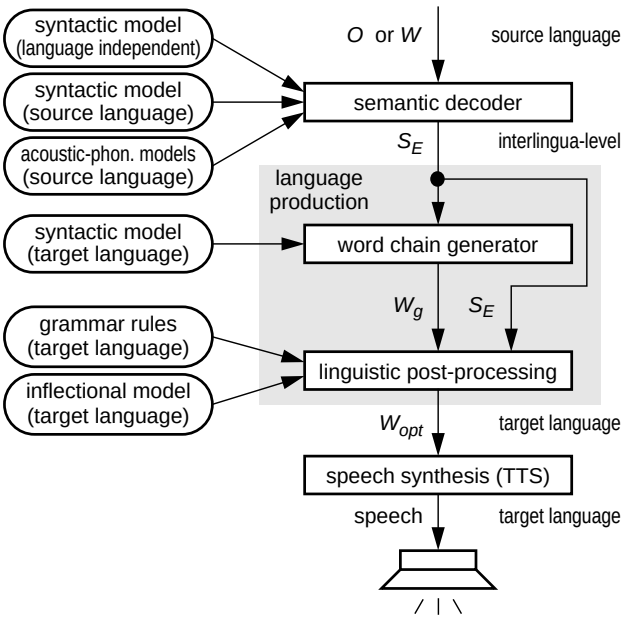

Figure 1: Overview of the automatic speech translation

- The semantic decoder converts either an observation sequence O (spoken utterance) or a word chain W (written text) of the source language into a corresponding semantic representation – in our approach the semantic structure S_E.

Under the assumption that a semantic structure is not language specific, we can use this semantic representation level as interlingua-level for the translation task – similar to the principle of the concept-based translation [4].

- The word chain generator converts a semantic structure S_E into a corresponding word chain W_g of the target language. Since the semantic structure S_E does not contain any grammatical information, the syntax of the word chain W_g may not be correct.
- The linguistic post-processing module converts the possibly grammatically wrong word chain W_g into a correct ('optimized') word chain W_{opt}.

These three components are described in detail in the following chapters. Furthermore, a speech synthesis module (text-to-speech, TTS) produces spoken output in the target language.

2. THE SEMANTIC STRUCTURE

In our approach, the semantic structure S (representing the semantic content) is a tree consisting of a finite number N of semantic units (simply called "semuns") s_n:

$$S = \{s_1, s_2, \ldots, s_n, \ldots, s_N\} \tag{1}$$

Each semun corresponds to exactly one significant word and not more than one insignificant word out of W. It expresses a small semantic partition of the utterance (i.e. the semantic contribution of the significant word), similar to "conceptual labels" [7] [8].

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 particular successor-semuns[1] $q_1[s_n], \ldots, q_X[s_n] \in \{s_2, \ldots, s_N, \text{blnk}\} \setminus \{s_n\}$:

$$s_n = \left(t[s_n], v[s_n], q_1[s_n], \ldots, q_X[s_n] \right), \; X \ge 1 \tag{2}$$

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

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 [9]. 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 $t[q_1[s_n]] = \text{blnk}$. A detailed description of the semantic structure can be found in [5].

As an example, fig. 2 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'):

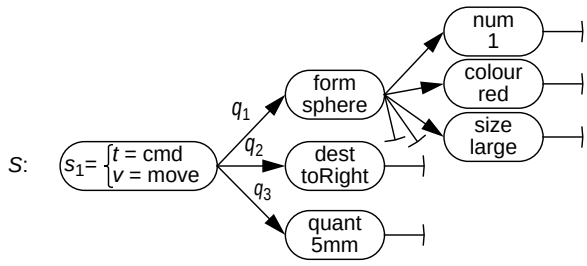

Figure 2: Semantic structure S in a graphic form

3. SEMANTIC DECODING

The semantic decoder converts a spoken utterance (given as observation sequence O) into its semantic representation (in our approach denoted as semantic structure S). From the set of all possible S, that one S_E has to be found which is most probable given the observation sequence O, i.e. which maximizes the a-posteriori-probability $P(S|O)$. The resulting term can be transformed using the Bayes formula.

$$S_E = \underset{S}{\operatorname{argmax}}\, P(S|O) = \underset{S}{\operatorname{argmax}} \frac{P(O|S) \cdot P(S)}{P(O)} \tag{3}$$

Since $P(O)$ is not relevant for maximizing, it can be neglected:

$$S_E = \underset{S}{\operatorname{argmax}}\, [P(O|S) \cdot P(S)] \tag{4}$$

Due to the high variety of S and O, it is not possible to estimate $P(O|S)$ directly. Therefore, additional representation levels are necessary. Clearly defined are the word chain W and the phoneme chain Ph, which can be used to calculate S_E:

$$S_E = \underset{S}{\operatorname{argmax}} \max_{W} \max_{Ph} [P(O|Ph) \cdot P(Ph|W) \cdot P(W|S) \cdot P(S)] \tag{5}$$

Eq. (5) can be implemented as 'top-down'-approach for finding that semantic structure S_E, which is the most likely combination of a semantic structure S, a word chain W, a phoneme chain Ph and the given observation sequence O. In the above equations, we assume statistical independence of all probabilities, which are stored in four stochastic knowledge bases (called "models"):

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

Phonetic and acoustic models are not necessary to decode written text, since in case of written input (i.e. word chain W) eq. (5) can be simplified to $S_E = \underset{S}{\operatorname{argmax}}\, [P(W|S) \cdot P(S)]$.

These probabilities have to be estimated by counting the occurring frequencies over a large set of training data, which are authentic limited-domain utterances, each represented by semantic structure, word chain, phoneme chain and observation sequence [6]. A detailed description of the semantic decoder, which is implemented as an incremental 'top-down'-parser combining a modified Earley-parsing [2] and a Viterbi-beam-search [15] algorithm, and the consistent integration of all stochastic knowledge can be found in [13].

4. WORD CHAIN GENERATOR

The word chain generator converts a semantic structure into the corresponding word chain of the target language. For this purpose, we make use of the generative power of our syntactic models by creating just the most likely word chain given a certain semantic structure. The stochastic process of originating word chains with the grammar described in [12] can be seen as a complex transition network similar to a hierarchic Hidden-Markov-Model [10]. Each semun corresponds to a "syntactic module" (SM), which is according to the given semantic structure S_E hierarchically connected with other SMs to a whole syntactic network. The transitions within this network affect the word alignment, the emissions affect the respective word choice. $P(W|S)$ is calculated as the product over all transition and emission probabilities along a certain path through the entire syntactic network:

$$P(W|S) = \prod_{\substack{\text{all transitions}\\ \text{along path}}} (\text{trans. prob.}) \cdot \prod_{\substack{\text{all states with}\\ \text{word emissions}\\ \text{along path}}} (\text{emiss. prob.}) \tag{6}$$

Unlike the word chain generator in the 'top-down' speech understanding (e.g. described in [1] or [11]), which has to produce many word chain hypotheses, this one delivers only the most probable word chain W_g given the semantic structure S_E. The concerning syntactic model considers that word chain W_g, which maximizes eq. (6):

$$W_g = \underset{W}{\operatorname{argmax}}\, P(W|S_E) \tag{7}$$

Fig. 3 depicts one selected path through such a syntactic network, consisting of four SMs:

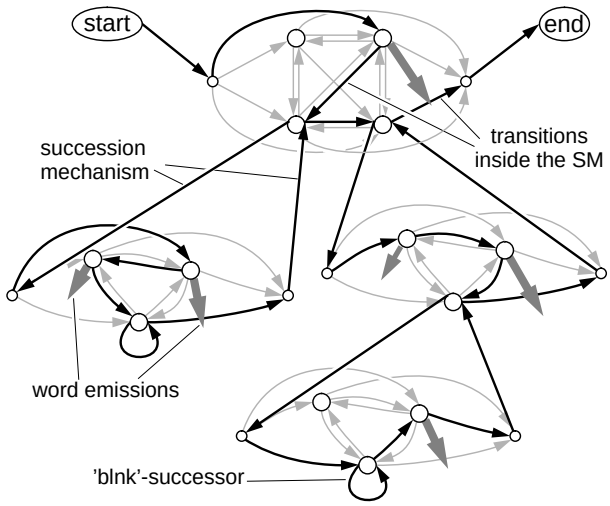

Figure 3: Origination of a word chain along a certain path by transitions and emissions within the syntactic network [12]

Since the semantic structure does not contain any grammatical information, case, number, and gender of words in W_g may be wrong. Nevertheless, W_g is usually comprehensible by humans. According to eq. (7), a word chain

W_g : 'erzeuge zwei rot kugel'

('create two red sphere') might be generated. In this case, the emission probabilities for the singular words 'rot' and 'kugel' is according to the syntactic model higher than those for the correct plural words 'rote' and 'kugeln'.

Because choice (i.e. the uninflected word) and alignment of the generated words in W_g are fixed by the transitions and the emissions of the syntactic model and cannot be re-changed in a later stage, it is advisable to use manually created instead of trained syntactic knowledge bases for the target language.

5. LINGUISTIC POST-PROCESSING

The post-processing module converts the grammatically wrong word chain W_g into a correct word chain W_{opt} by changing the inflection of some words using the following knowledge bases:

- The grammar rules specify the grammatical features (case, number, gender) of a certain word by the help of both the word chain W_g and the sem. structure S_E.
- The inflectional model delivers the correct inflection of a word given its grammatical features.

The semantic structure S_E is recursively examined semun by semun. If the affiliated word of the current semun is a noun, an adjective, or an article, the correct inflection has to be found according to the respective grammatical features. The gender is extracted depending on the emitted noun, but case and number are extracted depending only on the semantic structure S_E (which is a different and new approach compared to classic linguistics). As an example, we look at following word chain W_g and semantic structure S_E:

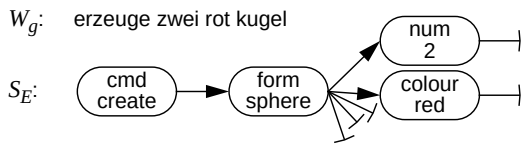

Figure 4: Word chain W_g and semantic structure S_E

- The semun '(cmd create)' corresponds to the word 'erzeuge'. As a verb, it remains unchanged.
- The semun '(form sphere)' corresponds to the word 'kugel'. As a noun, the grammatical features are determined as follows:

  Since the predecessor-semun is '(cmd create)', the case is accusative. Since the first successor-semun is '(num 2)', the number is plural. Since the emitted noun is 'kugel', the gender is feminine.

  The inflectional model delivers for the word 'kugel' with the grammatical features "accusative, plural, feminine" the correct word 'kugeln'.
- The semun '(num 2)' corresponds to the word 'zwei'. As a number, it remains unchanged.
- The semun '(colour red)' corresponds to the word 'rot'. As an adjective, the grammatical features are determined as follows:

  Since the predecessor-semun is '(form sphere)', case, number, and gender are identical to the corresponding word of the predecessor-semun.

  The inflectional model delivers for the word 'rot' with the grammatical features "accusative, plural, feminine" the correct word 'rote'.

After gone through the whole semantic structure, the optimized word chain W_{opt} with the correct inflections can be composed as

W_{opt} : 'erzeuge zwei rote kugeln'

('create two red spheres'). Note, that neither the choice nor the alignment of the words is changed during the post-processing.

6. PERFORMANCE RATES

6.1. Semantic Decoding

For spoken input, the grammar was trained with 1843 utterances within the 'graphic editor' domain. For testing the decoder, we used 100 utterances, which are a subset of the training set, to avoid out-of-vocabulary errors. The utterances have been collected by a 'Wizard-of-Oz' experiment with 33 speakers [6]. Each utterance is represented by the speech signal (16 kHz, 16 bit) and the associated semantic structure, which was used both for training and as reference for testing the decoder's semantic accuracy. In the tests, the semantic errors are determined as the percentage of wrongly assigned semantic structures. Fig. 5 shows the semantic decoding errors and the computation effort related to the beam width. It can be seen that pruning is necessary to prevent an uncontrollable increase of the

computations. A compromise between performance and computation effort results in a semantic error rate of approximately 30%.

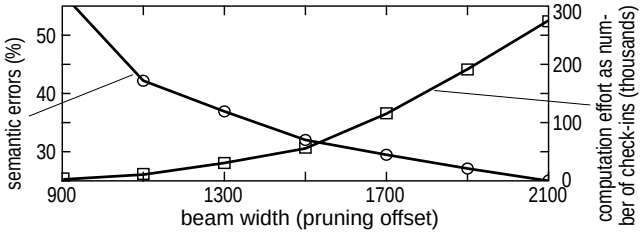

Figure 5: Semantic error rate for spoken input [13]

For written input, the wrong conversion of a word chain into a semantic structure amounts to only 0.2%, if the test set is a subset of the training material. If the test set is not included in the training, the rate increases to 22.8%, because of many out-of-vocabulary errors due to unknown words [12].

6.2. Language Production

The language production was tested with 307 real existing semantic structures within the 'graphic editor' domain producing four different target languages German, English, French, and Slovenian. We asked human subjects to judge the semantics (whether it is comprehensible or incomprehensible) and the syntax (whether it is correct, unusual, or wrong) for every optimized word chain W_{opt}.

target language	semantics	syntax	
	comprehensible	correct	unusual
German	95.9 %	84.4 %	5.2 %
English	94.8 %	81.1 %	10.4 %
French	93.8 %	82.4 %	9.8 %
Slovenian	88.3 %	82.1 %	8.5 %

Table 1: Language production performance

The above results with an average comprehensible semantics of 93.2% and an average not-wrong syntax of 91.0% confirm that the introduced translation approach based on the semantic structure can be an alternative to other much more complex rule-based and word-based approaches, if short sentences within a limited domain should be translated.

7. ACKNOWLEDGEMENTS

The authors want to thank Janez Kaiser (Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia) and Elisabeth Müller (Faculty of Romance Languages, Ludwig-Maximilians-University, Munich, Germany) for designing the Slovenian and French language production knowledge bases.

An online system for the automatic translation of German word chains into English or French word chains within the 'graphic editor' domain is running on WWW. If you try it, please be aware of out-of-vocabulary errors. The internet-address is as follows:

http://www.mmk.e-technik.tu-muenchen.de/~mue/nasgra/

8. 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), pp. 567-570
2. J. Earley: *An Efficient Context-Free Parsing Algorithm*, Comm. of the ACM, vol. 13 (1970), no. 2, pp. 94-102
3. V. M. Jiménez, A. Castellanos, E. Vidal: *Some Results with a Trainable Speech Translation and Understanding System*, Proc. ICASSP 1995 (Detroit, USA), pp. 113-116
4. L. Mayfield, M. Gavalda, W. Ward, A. Waibel: *Concept-based Speech Translation*, Proc. ICASSP 1995 (Detroit, USA), pp. 97-100
5. J. Müller, H. Stahl: *Die semantische Gliederung als adäquate semantische Repräsentationsebene für einen sprachverstehenden 'Grafikeditor'*, in L. Hitzenberger (ed.): "Angewandte Computerlinguistik", Georg Olms Publishing, Hildesheim, 1995, pp. 211-225 (in German)
6. J. Müller, H. Stahl: *Collecting and Analyzing Spoken Utterances for a Speech Controlled Application*, Proc. Eurospeech 1995 (Madrid, Spain), pp. 1437-1440
7. R. Pieraccini et al.: *A Speech Understanding System Based on Statistical Representation of Semantics*, Proc. ICASSP 1992 (San Francisco, USA), pp. I.193-I.196
8. R. Pieraccini, E. Levin, E. Vidal: *Learning how to Understand Language*, Proc. Eurospeech 1993 (Berlin, Germany), pp. 1407-1412
9. M. Pinkal: *Semantik*, in G. Görz (ed.): "Einführung in die künstliche Intelligenz", Addison-Wesley, Bonn, 1993, pp. 425-498 (in German)
10. L. R. Rabiner: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proc. IEEE, vol. 77 (1989), no. 2, pp. 257-286
11. 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-Communic.' 1994 (Maribor, Slovenia), pp. 16.1-16.9
12. H. Stahl, J. Müller: *A Stochastic Grammar for Isolated Representation of Syntactic and Semantic Knowledge*, Proc. Eurospeech 1995 (Madrid, Spain), pp. 551-554
13. H. Stahl, J. Müller, M. Lang: *An Efficient Top-Down Parsing Algorithm for Understanding Speech by Using Stochastic Syntactic and Semantic Models*, Proc. ICASSP 1996 (Atlanta, USA), pp. I.397-I.400
14. E. Vidal: *Language Learning, Understanding and Translation*, Proc. CRIM-FORWISS Workshop on "Progress and Prospects of Speech Research and Technology" (Munich, Germany), 1994, pp. 131-140
15. A.J. Viterbi: *Error Bounds for Convolutional Codes and an Asymptotical Optimal Decoding Algorithm*, IEEE Trans. Information Theory, vol. 61 (1973), pp. 268-278