STATISTICAL TRANSLATION OF SPOKEN DIALOGUES IN THE VERBMOBIL SYSTEM

H. Ney, F. J. Och, S. Vogel

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen, University of Technology
D-52056 Aachen, Germany

ABSTRACT

This paper gives an overview of our work on statistical machine translation in the framework of the VERBMOBIL project. The goal of the VERBMOBIL project is the translation of spoken dialogues in the domains of appointment scheduling and travel planning. Starting with the Bayes decision rule as in speech recognition, we show how the required probability distributions can be structured into three parts: the language model, the alignment model and the lexicon model. We describe the components of the system and report results on the VERBMOBIL task. The experience obtained in the VERBMOBIL project, in particular a large-scale end-to-end evaluation, showed that the statistical approach resulted in significantly lower error rates than three competing translation approaches: the sentence error rate was 29% in comparison with 52% to 62% for the other translation approaches.

1. INTRODUCTION

In comparison with written language, speech and especially spontaneous speech poses additional difficulties for the task of automatic translation. Typically, these difficulties are caused by errors of the recognition process, which is carried out before the translation process. As a result, the sentence to be translated is not necessarily well-formed from a syntactic point-of-view. Even without recognition errors, speech translation has to cope with a lack of conventional syntactic structures because the structures of spontaneous speech differ from that of written language.

The statistical approach shows the potential to tackle these problems for the following reasons. First, the statistical approach is able to avoid hard decisions at any level of the translation process. Second, for any source sentence, a translated sentence in the target language is guaranteed to be generated. In most cases, this will be hopefully a syntactically perfect sentence in the target language; but even if this is not the case, in most cases, the translated sentence will convey the meaning of the spoken sentence.

Whereas statistical modelling is widely used in speech recognition, there are so far only a few research groups that apply statistical modelling to language translation. The presentation here is based on work carried out in the framework of the EUTRANS project [1] and the VERBMOBIL project [21].

2. STATISTICAL DECISION THEORY AND LINGUISTICS

2.1. The Statistical Approach

The use of statistics in computational linguistics has been extremely controversial for more than three decades. The controversy is very well summarized by the statement of Chomsky in 1969 [6]:

> "It must be recognized that the notion of a 'probability of a sentence' is an entirely useless one, under any interpretation of this term".

This statement was considered to be true by the majority of experts from artificial intelligence and computational linguistics, and the concept of statistics was banned from computational linguistics for many years.

What is overlooked in this statement is the fact that, in an automatic system for speech recognition or text translation, we are faced with the problem of taking decisions. It is exactly here where statistical decision theory comes in. In speech recognition, the success of the statistical approach is based on the equation:

$$\begin{aligned}\text{Speech Recognition} \quad = \quad & \text{Acoustic–Linguistic Modelling} \\ & + \text{Statistical Decision Theory}\end{aligned}$$

Similarly, for machine translation, the statistical approach is expressed by the equation:

$$\begin{aligned}\text{Machine Translation} \quad = \quad & \text{Linguistic Modelling} \\ & + \text{Statistical Decision Theory}\end{aligned}$$

For the 'low-level' description of speech and image signals, it is widely accepted that the statistical framework allows an efficient coupling between the observations and the models, which is often described by the buzz word 'subsymbolic processing'. But there is another advantage in using probability distributions in that they offer an explicit formalism for expressing and combining hypothesis scores:

- The probabilities are directly used as scores: These scores are normalized, which is a desirable property: when increasing the score for a certain element in the set of all hypotheses, there must be one or several other elements whose scores are reduced at the same time.
- It is straightforward to combine scores: depending on the task, the probabilities are either multiplied or added.

- Weak and vague dependencies can be modelled easily. Especially in spoken and written natural language, there are nuances and shades that require 'grey levels' between 0 and 1.

Even if we think we can manage without statistics, we will need models which always have some free parameters. Then the question is how to train these free parameters. The obvious approach is to adjust these parameters in such a way that we get optimal results in terms of error rates or similar criteria on a representative sample. So we have made a complete cycle and have reached the starting point of the statistical modelling approach again!

When building an automatic system for speech or language, we should try to use as much prior knowledge as possible about the task under consideration. This knowledge is used to guide the modelling process and to enable improved generalization with respect to unseen data. Therefore in a good statistical modelling approach, we try to identify the common patterns underlying the observations, i.e. to capture dependencies between the data in order to avoid the pure 'black box' concept.

2.2. Bayes Decision Rule and System Architecture

In machine translation, the goal is the translation of a text given in a source language into a target language. We are given a source string $f_1^J = f_1...f_j...f_J$, which is to be translated into a target string $e_1^I = e_1...e_i...e_I$. In this article, the term *word* always refers to a *full-form* word. Among all possible target strings, we will choose the string with the highest probability which is given by Bayes decision rule [5]:

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{e_1^I} \{Pr(e_1^I|f_1^J)\} \\ &= \arg\max_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\} \quad . \end{aligned}$$

Here, $Pr(e_1^I)$ is the language model of the target language, and $Pr(f_1^J|e_1^I)$ is the string translation model. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. The overall architecture of the statistical translation approach is summarized in Figure 1.

In general, as shown in this figure, there may be additional transformations to make the translation task simpler for the algorithm. The transformations may range from the categorization of single words and word groups to more complex preprocessing steps that require some parsing of the source string. We have to keep in mind that in the search procedure both the language and the translation model are applied *after* the text transformation steps. However, to keep the notation simple, we will not make this explicit distinction in the subsequent exposition.

3. ALIGNMENT MODELLING

3.1. Concept

A key issue in modelling the string translation probability $Pr(f_1^J|e_1^I)$ is the question of how we define the correspondence between the words of the target sentence and the words of the source sentence. In typical cases, we can assume a sort of pairwise dependence by considering all word pairs (f_j, e_i) for a given sentence pair $(f_1^J; e_1^I)$. Here, we will

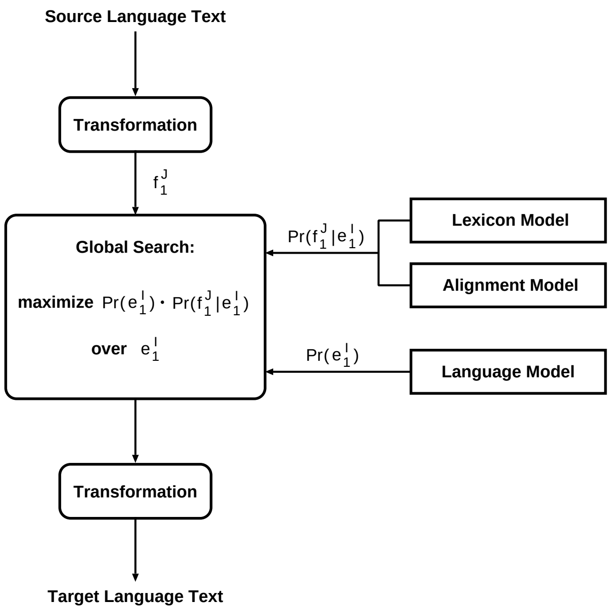

Figure 1. Architecture of the translation approach based on Bayes decision rule.

further constrain this model by assigning each source word to *exactly one* target word. Later, this requirement will be relaxed. Models describing these types of dependencies are referred to as *alignment models* [5, 20].

When aligning the words in parallel texts, we typically observe a strong localization effect. Figure 2 illustrates this effect for the language pair German–English. In many cases, although not always, there is an additional property: over large portions of the source string, the alignment is monotone.

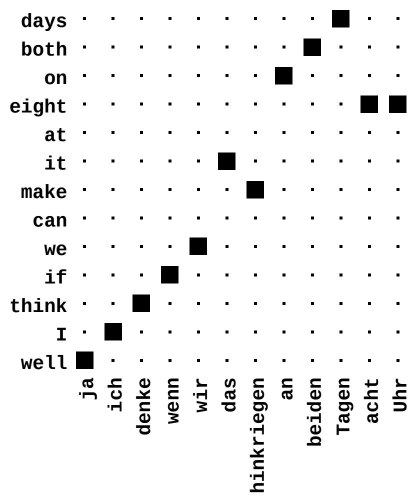

Figure 2. Word-to-word alignment.

3.2. Basic Models

To arrive at a quantitative specification, we define the alignment mapping: $j \to i = a_j$, which assigns a word f_j in position j to a word e_i in position $i = a_j$. We rewrite the probability for the translation model by introducing the 'hidden' alignments $a_1^J := a_1 ... a_j ... a_J$ for each sentence pair $(f_1^J; e_1^I)$. To structure this probability distribution, we factorize it over the positions in the source sentence and limit the alignment dependencies to a first-order dependence:

$$Pr(f_1^J|e_1^I) \quad = \quad p(J|I) \cdot \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I, J) \cdot p(f_j|e_{a_j})] \ .$$

Here, we have the following probability distributions:

- the sentence length probability: $p(J|I)$, which is included here for completeness, but can be omitted without loss of performance;
- the lexicon probability: $p(f|e)$;
- the alignment probability: $p(a_j|a_{j-1}, I, J)$.

By making the alignment probability $p(a_j|a_{j-1}, I, J)$ dependent on the jump width $a_j - a_{j-1}$ instead of the absolute positions a_j, we obtain the so-called homogeneous hidden Markov model, for short HMM [20].

We can also use a *zero-order* model $p(a_j|j, I, J)$, where there is only a dependence on the *absolute* position index j of the source string. This is the so-called model IBM-2 [5]. Assuming a uniform alignment probability $p(a_j|j, I, J) = 1/I$, we arrive at the so-called model IBM-1.

These models can be extended to allow for source words having no counterpart in the translation. Formally, this is incorporated into the alignment models by adding a so-called 'empty word' at position $i = 0$ to the target sentence and aligning all source words without a direct translation to this empty word.

In [5], more refined alignment models are introduced by using the concept of fertility. The idea is that often a word in the target language may be aligned to several words in the source language. This is the so-called model IBM-3. Using, in addition, first-order alignment probabilities along the positions of the source string leads us to model IBM-4. Although these models take one-to-many alignments explicitly into account, the lexicon probabilities $p(f|e)$ are still based on single words in each of the two languages.

In systematic experiments, it was found that the quality of the alignments determined from the bilingual training corpus has a direct effect on the translation quality [13].

3.3. Alignment Template Approach

A general shortcoming of the baseline alignment models is that they are mainly designed to model the lexicon dependences between single words. Therefore, we have extended the approach to handle word groups or phrases rather than single words as the basis for the alignment models [12]. In other words, a whole group of adjacent words in the source sentence may be aligned with a whole group of adjacent words in the target language. As a result, the context of words tends to be explicitly taken into account, and the differences in local word orders between source and target languages can be learned explicitly. Figure 3 shows some of the extracted alignment templates for a sentence pair from the VERBMOBIL training corpus. The training algorithm for the alignment templates extracts all phrase pairs which are aligned in the training corpus up to a maximum length of 7 words. To improve the generalization capability of the

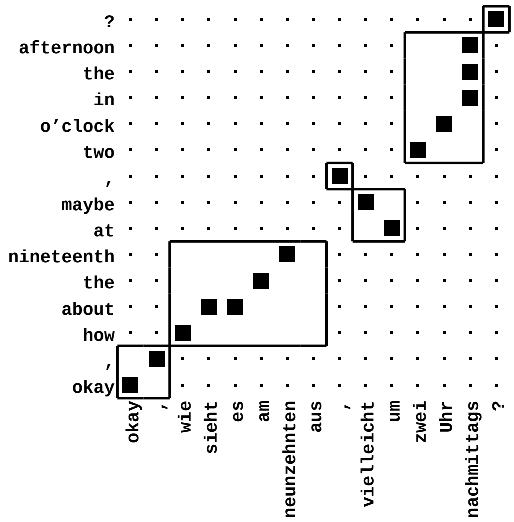

Figure 3. Example of a word alignment and of extracted alignment templates.

alignment templates, the templates are determined for word classes rather than words directly. These word classes are determined by an automatic clustering procedure [11].

4. SEARCH

The task of the search algorithm is to generate the most likely target sentence e_1^I of unknown length I for an observed source sentence f_1^J. The search must make use of all three knowledge sources as illustrated by Figure 4: the alignment model, the lexicon model and the language model. All three of them must contribute in the final decision about the words in the target language.

To illustrate the specific details of the search problem, we use the *inverted* alignment: $i \rightarrow j = b_i$, which is a mapping from *target* to *source* positions rather the other way round. We replace the sum over all alignments by the best alignment, which is referred to as maximum approximation in speech recognition. Using a bigram language model $p(e_i|e_{i-1})$, we obtain the following search criterion:

$$\max_{b_1^I, e_1^I} \prod_{i=1}^{I} [p(e_i|e_{i-1}) \cdot p(b_i|b_{i-1}, J) \cdot p(f_{b_i}|e_i)]$$

Considering this criterion, we can see that we can build up hypotheses of partial target sentences in a *bottom-to-top* strategy over the positions i of the target sentence e_1^i as illustrated in Figure 5. An important constraint for the alignment is that *all* positions of the source sentence should be covered exactly *once*. This constraint is similar to that of the travelling salesman problem where each city has to be visited exactly once. Details on various search strategies can be found in [8, 9, 18].

In order to take long context dependences into account, we use a class-based five-gram language model with backing-off. Beam-search is used to handle the huge search space. To normalize the costs of partial hypotheses covering different parts of the input sentence, an (optimistic) estimation of the remaining cost is added to the current

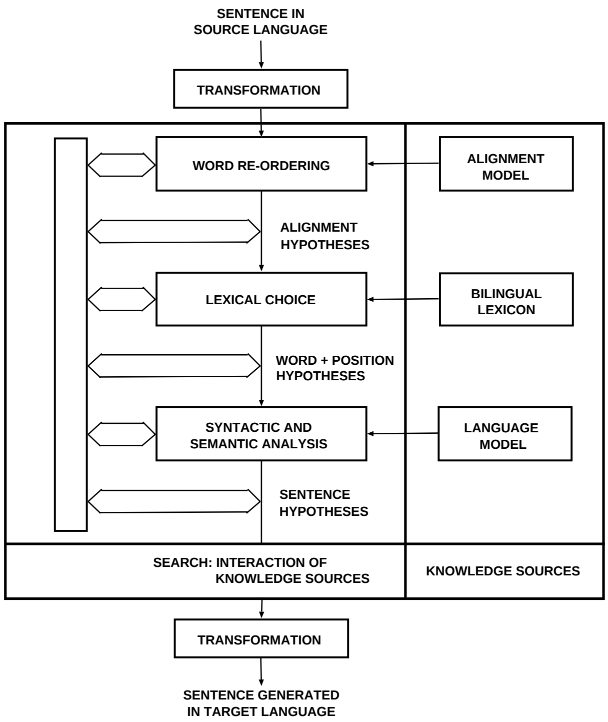

Figure 4. Illustration of search in statistical translation.

accumulated cost as follows. For each word in the source sentence, a lower bound on its translation cost is determined beforehand. Using this lower bound, it is possible to achieve an efficient estimation of the remaining cost.

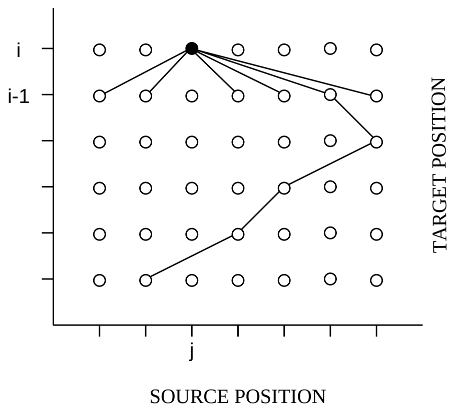

Figure 5. Illustration of bottom-to-top search.

5. EXPERIMENTAL RESULTS

5.1. The Task and the Corpus

Within the VERBMOBIL project, spoken dialogs were recorded. These dialogs were manually transcribed and later manually translated by VERBMOBIL partners (Hildesheim for Phase I and Tübingen for Phase II). Since different human translators were involved, there is great variability in the translations.

Each of these so-called dialog turns may consist of several sentences spoken by the same speaker and is sometimes rather long. As a result, there is no one-to-one correspondence between source and target sentences. To achieve a one-to-one correspondence, the dialog turns are split into shorter segments using punctuation marks as potential split points. Since the punctuation marks in source and target sentences are not necessarily identical, a dynamic programming approach is used to find the optimal segmentation points. The number of segments in the source sentence and in the test sentence can be different. The segmentation is scored using a word-based alignment model, and the segmentation with the best score is selected. This segmented corpus is the starting point for the training of translation and language models. Alignment models of increasing complexity are trained on this bilingual corpus [13].

A standard vocabulary had been defined for the various speech recognizers used in VERBMOBIL. However, not all words of this vocabulary were observed in the training corpus. Therefore, the translation vocabulary was extended semi-automatically using an online bilingual lexicon available on the web. The resulting lexicon contained not only word-word entries, but also multi-word translations, especially for the large number of German compound words. To counteract the sparseness of the training data, a couple of straightforward rule-based preprocessing steps are applied:

- categorization of proper names for persons and cities,
- normalization of:
  - numbers,
  - time and date phrases,
  - spelling: `don't` $\rightarrow$ `do not`,...
- splitting of
  German compound words.

Table 1 gives the characteristics of the training corpus and the lexicon. The 58 000 sentence pairs comprise about half a million running words for each language of the bilingual training corpus. The vocabulary size given is the number of full word forms seen in that corpus including the punctuation marks. Notice the large number of word types seen only once. The extended vocabulary is the vocabulary after adding the manual bilingual lexicon.

Table 1. Bilingual training corpus (PM = punctuation mark) and lexica.

		German	English
Training Text	Sentences	58 332	
	Words (+PMs)	519 523	549 921
	Vocabulary	7 940	4 673
	Singletons	44.8%	37.6%
Recognition	Vocabulary	10 157	6 871
Manual Lexicon	Entry Pairs	12 779	
	Ext. Vocab.	11 501	6 867

5.2. Disambiguation Examples

In the statistical translation approach as we have presented it, no explicit word sense disambiguation is performed. However, a kind of implicit disambiguation is possible due to the context information of the alignment templates and the language model as shown by the examples in Table 2. The first two groups of sentences contain the verbs 'gehen' and 'annehmen' which have different translations, some of

Table 2. Disambiguation examples (* = using morpho-syntactic analysis.)

Ambiguous Word	Text Input	Translation
gehen	Wir gehen ins Theater.	We will go to the theater.
	Mir geht es gut.	I am fine.
	Es geht um Geld.	It is about money.
	Geht es bei Ihnen am Montag?	Is it possible for you on Monday?
	Das Treffen geht bis 5 Uhr.	The meeting is to five.
annehmen	Wir sollten das Angebot annehmen.	We should accept that offer.
	Ich nehme das Schlimmste an.	I will assume the worst.*
vor	Wir treffen uns vor dem Frühstück.	We meet before the breakfast.
	Wir treffen uns vor dem Hotel.	We will meet in front of the hotel.

which are rather collocational. The correct translation is only possible by taking the whole sentence into account. Some improvement can be achieved by applying morpho-syntactic analysis, e.g handling of the separated verb prefixes in German [10].

The last two sentences show the implicit disambiguation of the temporal and spatial sense for the German preposition 'vor'. Although the system has not been tailored to handle such types of disambiguation, the translated sentences are all acceptable, apart from the sentence: `The meeting is to five`.

5.3. Integration into VERBMOBIL Prototype System

The statistical approach to machine translation is embodied in the *stattrans* module which is integrated into the VERBMOBIL prototype system. The implementation supports the translation directions from German to English and from English to German. In regular processing mode, the *stattrans* module receives its input from the *repair* module [16]. At that time, the word lattices and best hypotheses from the speech recognition systems have already been prosodically annotated, i.e. information about prosodic segment boundaries, sentence mode and accentuated syllables are added to each edge in the word lattice [3]. The translation is performed on the single best sentence hypothesis of the recognizer.

The prosodic boundaries and the sentence mode information are utilized by the *stattrans* module as follows. If there is a major phrase boundary, a full stop or question mark is inserted into the word sequence, depending on the sentence mode as indicated by the *prosody* module. Additional commas are inserted for other types of segment boundaries. The *prosody* module calculates probabilities for segment boundaries, and thresholds are used to decide if the sentence marks are to be inserted. These thresholds have been selected in such a way that, on the average, for each dialog turn, a good segmentation is obtained. The segment boundaries restrict possible word reordering between source and target language. This not only improves translation quality, but also restricts the search space and thereby speeds up the translation process.

5.4. Large-Scale End-to-End Evaluation

During the progress of the VERBMOBIL project, different variants of statistical translation have been implemented, and experimental tests have been performed for both text and speech input [8, 12].

Whereas these tests were important for the optimization and tuning of the system, the most important evaluation was the final evaluation of the VERBMOBIL prototype in spring 2000. This end-to-end evaluation of the VERBMOBIL system was performed at the University of Hamburg [17]. In each session of this evaluation, two native speakers conducted a dialog. They did not have any direct contact and could only interact by speaking and listening to the VERBMOBIL system.

Three other translation approaches had been integrated into the VERBMOBIL prototype system:

- a classical transfer approach [4, 7, 19], which is based on a manually designed analysis grammar, a set of transfer rules, and a generation grammar,
- a dialog-act based approach [14], which amounts to a sort of *slot filling* by classifying each sentence into one out of a small number of possible sentence patterns and filling in the slot values,
- an example-based approach [2], where a sort of nearest neighbour concept is applied to the set of bilingual training sentence pairs after suitable preprocessing.

In the final end-to-end evaluation human evaluators judged the translation quality for each of the four translation results using the following criterion:

Is the sentence approximatively correct: yes/no?

The evaluators were asked to pay particular attention to the semantic information (e.g. date and place of meeting, etc) contained in the translation. A missing translation as it may happen for the transfer approach or other approaches was counted as wrong translation. The evaluation was based on 5069 dialog turns for the translation from German to English and on 4136 dialog turns for the translation from English to German. The speech recognizers used had a word error rate of about 25%. The sentence error rates are summarized in Table 3. As we can see, the error rates for the statistical approach are smaller by a factor of about 2 in comparison with the other approaches.

Table 3. Sentence error rates of end-to-end evaluation (speech recognizer with WER=25%).

Translation Method	Error [%]
Semantic Transfer	62
Dialog Act Based	60
Example Based	52
Statistical	29

In agreement with other evaluation experiments, these experiments show that the statistical modelling approach may be comparable to or better than the conventional rule-based approach. In particular, the statistical approach seems to have the advantage if robustness is important, e.g. when the input string is not grammatically correct or when it is corrupted by recognition errors.

Although both text and speech input are translated with good quality on the average, there are examples where the syntactic structure of the produced sentence is not correct. Some of these syntactic errors are related to long range dependencies and syntactic structures that are not captured by the m-gram language model used. To cope with these problems, morpho-syntactic analysis [10] and grammar-based language models [15] are currently being studied.

6. SUMMARY

In this paper, we have given an overview of the statistical approach to machine translation and especially its implementation in the VERBMOBIL prototype system. The statistical system has been trained on about 500 000 running words from a bilingual German–English corpus. Translations are performed for both directions, i.e. from German to English and from English to German. Comparative evaluations with other translation approaches of the VERBMOBIL prototype system show that the statistical translation is superior, especially in the presence of speech input and ungrammatical input.

Acknowledgment

The work reported here has been supported partly by the VERBMOBIL project (contract number 01 IV 701 T4) by the German Federal Ministry of Education, Science, Research and Technology and as part of the EUTRANS project (ESPRIT project number 30268) by the European Community.

REFERENCES

[1] J. C. Amengual, J. M. Benedí, F. Casacuberta et al.: EuTrans – Example-based language translation systems. Technical Report, First Year Progress Report, European Union, Project Number 30268, July 1998.

[2] M. Auerswald: Example-based machine translation with templates. In [21], pp. 418–427.

[3] A. Batliner, J. Buckow, H. Niemann, E. Nöth, V. Warnke: The prosody module. In [21], pp. 106–121.

[4] T. Becker, A. Kilger, P. Lopez, P. Poller: The Verbmobil generation component VM-GECO. In [21], pp. 481–496.

[5] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[6] N. Chomsky: "Quine's Empirical Assumptions", in D. Davidson, J. Hintikka (eds.): *Words and objections. Essays on the work of W. V. Quine*, Reidel, Dordrecht, The Netherlands, 1969.

[7] M. C. Emele, M. Dorna, A. Lüdeling, H. Zinsmeister, C. Rohrer: Semantic-based transfer. In [21], pp. 359–376.

[8] H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, S. Vogel: Algorithms for statistical translation of spoken language. *IEEE Trans. on Speech and Audio Processing* Vol. 8, No. 1, pp. 24–36, Jan. 2000.

[9] S. Nießen, S. Vogel, H. Ney, C. Tillmann: A DP based search algorithm for statistical machine translation. *COLING–ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics*, pp. 960–967, Montreal, Canada, Aug. 1998.

[10] S. Nießen, H. Ney: Improving SMT quality with morpho-syntactic analysis. *18th Int. Conf. on Computational Linguistics*, pp. 1081-1085, Saarbrücken, Germany, July 2000.

[11] F. J. Och: An efficient method to determine bilingual word classes. *9th Conf. of the European Chapter of the Association for Computational Linguistics*, pp. 71–76, Bergen, Norway, June 1999.

[12] F. J. Och, C. Tillmann, H. Ney: Improved alignment models for statistical machine translation. *Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, 20–28, University of Maryland, College Park, MD, June 1999.

[13] F. J. Och, H. Ney: A comparison of alignment models for statistical machine translation. *18th Int. Conf. on Computational Linguistics*, pp. 1086-1090, Saarbrücken, Germany, July 2000.

[14] N. Reithinger, R. Engel: Robust content extraction for translation and dialog processing. In [21], pp. 428–437.

[15] H. Sawaf, K. Schütz, H. Ney: On the use of grammar based language models for statistical machine translation. *6th Int. Workshop on Parsing Technologies*, pp. 231–241, Trento, Italy, Feb. 2000.

[16] J. Spilker, M. Klarner, G. Görz: Processing self-corrections in a speech-to-speech system. In [21], pp. 131–140.

[17] L. Tessiore, W. v. Hahn: Functional validation of a machine translation system: Verbmobil. In [21], pp. 611–631.

[18] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga: A DP-based search using monotone alignments in statistical translation. *35th Annual Conf. of the Association for Computational Linguistics*, pp. 289–296, Madrid, Spain, July 1997.

[19] H. Uszkoreit, D. Flickinger, W. Kasper, I. A. Sag: Deep linguistic analysis with HPSG. In [21], pp. 216–263.

[20] S. Vogel, H. Ney, C. Tillmann: HMM-based word alignment in statistical translation. *16th Int. Conf. on Computational Linguistics*, pp. 836–841, Copenhagen, Denmark, August 1996.

[21] W. Wahlster (Ed.): *Verbmobil: Foundations of speech-to-speech translations.* Springer-Verlag, Berlin, 2000.