Improving SMT quality with morpho-syntactic analysis

Sonja Nießen and ***Hermann Ney***
Lehrstuhl für Informatik VI
Computer Science Department
RWTH – University of Technology Aachen
D-52056 Aachen, Germany
Email: niessen@informatik.rwth-aachen.de

Abstract

In the framework of statistical machine translation (SMT), correspondences between the words in the source and the target language are learned from bilingual corpora on the basis of so-called alignment models. Many of the statistical systems use little or no linguistic knowledge to structure the underlying models. In this paper we argue that training data is typically not large enough to sufficiently represent the range of different phenomena in natural languages and that SMT can take advantage of the explicit introduction of some knowledge about the languages under consideration. The improvement of the translation results is demonstrated on two different German-English corpora.

1 Introduction

In this paper, we address the question of how morphological and syntactic analysis can help statistical machine translation (SMT). In our approach, we introduce several transformations to the source string (in our experiments the source language is German) to demonstrate how linguistic knowledge can improve translation results especially in the cases where the token-type ratio (number of training words versus number of vocabulary entries) is unfavorable.

After reviewing the statistical approach to machine translation, we first explain our motivation for examining additional knowledge sources. We then present our approach in detail. Experimental results on two bilingual German-English tasks are reported, namely the VERBMOBIL and the EUTRANS task. Finally, we give an outlook on our future work.

2 Statistical Machine Translation

The goal of the translation process in statistical machine translation can be formulated as follows: A source language string $f_1^J = f_1 \ldots f_J$ is to be translated into a target language string $e_1^I = e_1 \ldots e_I$. In the experiments reported in this paper, the source language is German and the target language is English. Every English string is considered as a possible translation for the input. If we assign a probability $Pr(e_1^I|f_1^J)$ to each pair of strings (e_1^I, f_1^J), then according to Bayes' decision rule, we have to choose the English string that maximizes the product of the English language model $Pr(e_1^I)$ and the string translation model $Pr(f_1^J|e_1^I)$.

Many existing systems for SMT (Wang and Waibel, 1997; Nießen et al., 1998; Och and Weber, 1998) make use of a special way of structuring the string translation model (Brown et al., 1993): The correspondence between the words in the source and the target string is described by alignments that assign one target word position to each source word position. The probability of a certain English word to occur in the target string is assumed to depend basically only on the source word aligned to it. It is clear that this assumption is not always valid for the translation of natural languages. It turns out that even those approaches that relax the word-by-word assumption like (Och et al., 1999) have problems with many phenomena typical of natural languages in general and German in particular like

- idiomatic expressions;
- compound words that have to be translated by more than one word;
- long range dependencies like prefixes of verbs placed at the end of the sentence;
- ambiguous words with different meanings dependent on the context.

The parameters of the statistical knowledge sources mentioned above are trained on bilingual corpora. Bearing in mind that more than 40% of the word forms have only been seen once in training (see Tables 1 and 4), it is obvious that the phenomena listed above can hardly be learned adequately from the data and that the explicit introduction of linguistic knowledge is expected to improve translation quality.

The overall architecture of the statistical translation approach is depicted in Figure 1. In this figure we already anticipate the fact that we will transform the source strings in a certain manner. If necessary we can also apply the inverse of these transformations on the produced output strings. In Section 3 we explain in detail which kinds of transformations we apply.

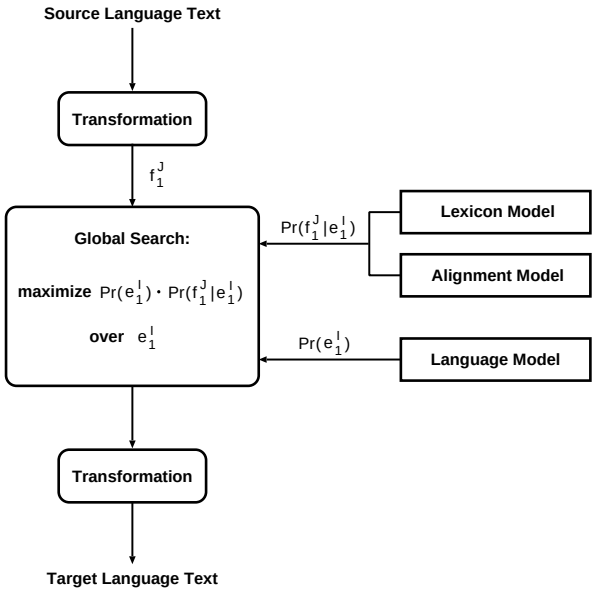

Figure 1: Architecture of the translation approach based on Bayes' decision rule.

3 Analysis and Transformation of the Input

As already pointed out, we used the method of transforming the input string in our experiments. The advantage of this approach is that existing training and search procedures did not have to be adapted to new models incorporating the information under consideration. On the other hand, it would be more elegant to leave the decision between different readings, for instance, to the overall decision process in search. The transformation method however is more adequate for the preliminary identification of those phenomena relevant for improving the translation results.

3.1 Analysis

We used GERTWOL, a German Morphological Analyser (Haapalainen and Majorin, 1995) and the Constraint Grammar Parser for German GERCG for lexical analysis and morphological and syntactic disambiguation. For a description of the Constraint Grammar approach we refer the reader to (Karlsson, 1990). Some preprocessing was necessary to meet the input format requirements of the tools. In the cases where the tools returned more than one reading, either simple heuristics based on domain specific preference rules where applied or a more general, non-ambiguous analysis was used.

In the following subsections we list some transformations we have tested.

3.2 Separated German Verbprefixes

Some verbs in German consist of a main part and a detachable prefix which can be shifted to the end of the clause, e.g. "losfahren" ("to leave") in the sentence "Ich fahre morgen los.". We extracted all word forms of separable verbs from the training corpus. The resulting list contains entries of the form `prefix|main`. The entry "los|fahre" indicates, for example, that the prefix "los" can be detached from the word form "fahre". In all clauses containing a word matching a main part and a word matching the corresponding prefix part occuring at the end of the clause, the prefix is prepended to the beginning of the main part, as in "Ich losfahre morgen."

3.3 German Compound Words

German compound words pose special problems to the robustness of a translation method, because the word itself must be represented in the training data: the occurence of each of the components is not enough. The word "Früchtetee" for example can not be translated although its components "Früchte" and "Tee" appear in the training set of EuTrans. Besides, even if the compound occurs in training, the training algorithm may not be capable of translating it properly as *two* words (in the mentioned case the words "fruit" and "tea") due to the word alignment assumption mentioned in Section 2. We

therefore split the compound words into their components.

3.4 Annotation with POS Tags

One way of helping the disambiguation of ambiguous words is to annotate them with their part of speech (POS) information. We chose the following very frequent short words that often caused errors in translation for VERBMOBIL:

"aber" can be adverb or conjunction.

"zu" can be adverb, preposition, separated verb prefix or infinitive marker.

"der", "die" and "das" can be definite articles or pronouns.

The difficulties due to these ambiguities are illustrated by the following examples: The sentence "Das würde mir sehr gut passen." is often translated by "*The* would suit me very well." instead of "*That* would suit me very well." and "Das war zu schnell." is translated by "That was *to* fast." instead of "That was *too* fast.".

We appended the POS tag in training and test corpus for the VERBMOBIL task (see 4.1).

3.5 Merging Phrases

Some multi-word phrases as a whole represent a distinct syntactic role in the sentence. The phrase "irgend etwas" ("anything") for example may form either an indefinite determiner or an indefinite pronoun. Like 21 other multi-word phrases "irgend-etwas" is merged in order to form one single vocabulary entry.

3.6 Treatment of Unseen Words

For statistical machine translation it is difficult to handle words not seen in training. For unknown proper names, it is normally correct to place the word unchanged into the translation. We have been working on the treatment of unknown words of other types. As already mentioned in Section 3.3, the splitting of compound words can reduce the number of unknown German words.

In addition, we have examined methods of replacing a word fullform by a more abstract word form and check whether this form is known and can be translated. The translation of the simplified word form is generally not the precise translation of the original one, but sometimes the intended semantics is conveyed, e.g.:

"kaltes" is an adjective in the singular neuter form and can be transformed to the less specific form "kalt" ("cold").

"Jahre" ("years") can be replaced by the singular form "Jahr".

"beneidest" ("to envy" in first person singular): if the infinitive form "beneiden" is not known, it might help just to remove the leading particle "be".

4 Translation Results

We use the SSER (subjective sentence error rate) (Nießen et al., 2000) as evaluation criterion: Each translated sentence is judged by a human examiner according to an error scale from 0.0 (semantically and syntactically correct) to 1.0 (completely wrong).

4.1 Translation Results for VERBMOBIL

The VERBMOBIL corpus consists of spontaneously spoken dialogs in the appointment scheduling domain (Wahlster, 1993). German sentences are translated into English. The output of the speech recognizer (for example the single-best hypothesis) is used as input to the translation modules. For research purposes the original text spoken by the users can be presented to the translation system to evaluate the MT component seperately from the recognizer.

The training set consists of 45 680 sentence pairs. Testing was carried out on a seperate set of 147 sentences that do not contain any unseen words. In Table 1 the characteristics of the training sets are summarized for the original corpus and after the application of the described transformations on the German part of the corpus. The table shows that on this corpus the splitting of compounds improves the token-type ratio from 59.7 to 65.2, but the number of singletons (words seen only once in training) does not go down by more than 2.8%. The other transformations (prepending separated verb prefixes "pref"; annotation with POS tags "pos"; merging of phrases "merge") do not affect these corpus statistics much.

The translation performance results are given in Table 2 for translation of text and in Table 3 for translation of the single-best hypothesis given by a speech recognizer (accuracy 69%).

For both cases, translation on text and on speech input, splitting compound words does

Table 1: Corpus statistics: VERBMOBIL training ("baseline"=no preprocessing).

preprocessing	no. of tokens	no. of types	single-tons
English	465 143	4 382	37.6%
German			
baseline	437 968	7 335	44.8%
verb prefixes	435 686	7 370	44.3%
split compounds	442 938	6 794	42.0%
pos	437 972	7 344	44.8%
pos+merge	437 330	7 363	44.7%
pos+merge+pref	435 055	7 397	44.2%

not improve translation quality, but it is not harmful either. The treatment of separable prefixes helps as does annotating some words with part of speech information. Merging of phrases does not improve the quality much further. The best translations were achieved with the combination of POS-annotation, phrase merging and prepending separated verb prefixes. This holds for both translation of text and of speech input.

Table 2: Results on VERBMOBIL text input.

preprocessing	SSER [%]
baseline	20.3
verb prefixes	19.4
split compounds	20.3
pos	19.7
pos+merge	19.5
pos+merge+pref	18.0

The fact that these hard-coded transformations are not only helpful on text input, but also on speech input is quite encouraging. As an example makes clear this cannot be taken for granted: The test sentence "Dann fahren wir dann los." is recognized as "Dann fahren wir dann uns." and the fact that separable verbs do not occur in their separated form in the training data is unfavorable in this case. The figures show that in general the speech recognizer output contains enough information for helpful preprocessing.

Table 3: Results on VERBMOBIL speech input.

preprocessing	SSER [%]
baseline	43.4
verb prefixes	41.8
split compounds	43.1
split+pref	42.3
pos+merge+pref	41.1

4.2 Translation Results for EUTRANS

The EUTRANS corpus consists of different types of German–English texts belonging to the tourism domain: web pages of hotels, touristic brochures and business correspondence. The string translation and language model parameters were trained on 27 028 sentence pairs. The 200 test sentences contain 150 words never seen in training.

Table 4 summarizes the corpus statistics of the training set for the original corpus, after splitting of compound words and after additional prepending of seperated verb prefixes ("split+prefixes"). The splitting of compounds improves the token-type ratio from 8.6 to 12.3 and the number of words seen only once in training reduces by 8.9%.

Table 4: Corpus statistics: EUTRANS.

preprocessing	no. of tokens	no. of types	single-tons
English	562 264	33 823	47.1%
German			
baseline	499 217	58 317	58.9%
split compounds	535 505	43 405	50.0%
split+prefixes	534 676	43 407	49.8%

The number of words in the test sentences never seen in training reduces from 150 to 81 by compound splitting and can further be reduced to 69 by replacing the unknown word forms by more general forms. 80 unknown words are encountered when verb prefixes are treated in addition to compound splitting.

Experiments for POS-annotation have not been performed on this corpus because no small set of ambiguous words causing many of the

translation errors on this task can be identified: Compared to the VERBMOBIL task, this corpus is less homogeneous. Merging of phrases did not help much on VERBMOBIL and is therefore not tested here.

Table 5 shows that the splitting of compound words yields an improvement in the subjective sentence error rate of 4.5% and the treatment of unknown words ("unk") improves the translation quality by an additional 1%. Treating separable verb prefixes in addition to splitting compounds gives the best result so far with an improvement of 7.1% absolute compared to the baseline.

Table 5: Results on EUTRANS.

preprocessing	SSER [%]
baseline	57.4
split compounds	52.9
split+unk	51.8
split+prefixes	50.3

5 Conclusion and Future Work

In this paper, we have presented some methods of providing morphological and syntactic information for improving the performance of statistical machine translation. First experiments prove their general applicability to realistic and complex tasks such as spontaneously spoken dialogs.

We are planning to integrate the approach into the search process. We are also working on language models and translation models that use morphological categories for smoothing in the case of unseen events.

Acknowledgement. This work was partly supported by the German Federal Ministry of Education, Science, Research and Technology under the Contract Number 01 IV 701 T4 (VERBMOBIL) and as part of the EUTRANS project by the European Community (ESPRIT project number 30268).

The authors would like to thank Gregor Leusch for his support in implementation.

References

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Mariikka Haapalainen and Ari Majorin. 1995. GERTWOL und Morphologische Disambiguierung für das Deutsche. URL: www.lingsoft.fi/doc/gercg/NODALIDA-poster.html.

Fred Karlsson. 1990. Constraint Grammar as a Framework for Parsing Running Text. In *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 168–173, Helsinki, Finland.

Sonja Nießen, Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1998. A DP based Search Algorithm for Statistical Machine Translation. In *Proceedings of the 36th Annual Conference of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 960–967, Montréal, P.Q., Canada, August.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 39–45, Athens, Greece, May.

Franz Josef Och and Hans Weber. 1998. Improving Statistical Natural Language Translation with Categories and Rules. In *Proceedings of the 36th Annual Conference of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 985–989, Montréal, P.Q., Canada, August.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, Maryland, June.

Wolfgang Wahlster. 1993. Verbmobil: Translation of Face-to-Face Dialogs. In *Proceedings of the MT Summit IV*, pages 127–135, Kobe, Japan.

Ye-Yi Wang and Alex Waibel. 1997. Decoding Algorithm in Statistical Translation. In *Proceedings of the ACL/EACL '97, Madrid, Spain*, pages 366–372, July.