Statistical Machine Translation of German Compound Words

Maja Popović, Daniel Stein, Hermann Ney

Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University
Ahornstrasse 55
52056 Aachen, Germany
{popovic,stein,ney}@informatik.rwth-aachen.de

Abstract. German compound words pose special problems to statistical machine translation systems: the occurence of each of the components in the training data is not sufficient for successful translation. Even if the compound itself has been seen during training, the system may not be capable of translating it properly into two or more words. If German is the target language, the system might generate only separated components or may not be capable of choosing the correct compound. In this work, we investigate and compare different strategies for the treatment of German compound words in statistical machine translation systems. For translation from German, we compare linguistic-based and corpus-based compound splitting. For translation into German, we investigate splitting and rejoining German compounds, as well as joining English potential components. Additionaly, we investigate word alignments enhanced with knowledge about the splitting points of German compounds. The translation quality is consistently improved by all methods for both translation directions.

1 Introduction

The goal of statistical machine translation is to translate an input word sequence in the source language into a target language word sequence. Given the source language sequence, we should choose the target language sequence which maximises the posterior probability. The translation system used in this work models this posterior probability directly as a log-linear combination of seven different models. The most important ones are phrase-based models in both directions. Additionaly, phrase level IBM1 models in both directions, a language model of the target language, as well as phrase penalty and word penalty are used. For detailed description of the system see [7, 8].

In order to improve the translation process, it is possible to perform preprocessing steps based on morphological and/or syntactic knowlegde in both the source and/or target language sequence. If necessary, after the translation the inverse transformations are applied to the generated target sequence.

In this work, we investigate and compare strategies for treatment of German compound words. For translation from German, we compare linguistic-based

and corpus-based approaches for splitting compounds in the source language. For translation into German, we explore two possibilities for improving the translation quality: splitting and rejoining German compounds and joining English words. Additionaly, we investigate how much the translation quality can be improved by incorporating knowledge about compound splitting points into the word alignments. This method is applied for both translation directions.

Related Work:

Several publications adress the problem of German compound words in statistical machine translation.

In [3], a morpho-syntactic analyser is used to split German compounds and improve the quality of the generated English output.

Corpus-based splitting for the same translation direction has been proposed in [1]. They compare several corpus-based methods and report that the one based on word frequencies yields the best translation improvements.

In this work, we compare these two methods on the European Parliament corpus. We propose several methods for treating German compounds when German is the target language. This problem has not be investigated yet to the best of our knowledge.

Some publications have proposed the use of morpho-syntactic knowledge for improving statistical alignment quality, for example [5, 6]. However, introducing knowledge about compound words has not been investigated so far.

In our work, we investigate the effects of introducing information about German compound words into the word alignments.

2 Treatment of German Compound Words

Compounding of words is common in many languages (German, Dutch, Finnish, etc.). Compound words are created by joining an arbitrary number of existing words together, and this can lead to a large increase of the vocabulary size, and thus also to sparse data problems. Therefore the problem of compound words poses challenges for many NLP applications. In this work, we investigate and compare different methods for treating German compound words in order to improve the quality of statistical machine translation both from German and into German.

2.1 Translation from German into English

For translation from German into English, the lingustic-based method proposed in [3] and the corpus-based method proposed in [1] are used in order to compare two approaches. For the linguistic-based splitting we used the Constraint Grammar Parser for German (GERCG) as described in [3]. For the corpus-based splitting we used the frequency-based method described in [1]:

- each capitalised word which consists of two or more words occuring in the training vocabulary is considered as a compound word

– for each compound word:
  - the frequency of the compound itself $N(w)$ and the frequencies of its components $N(w_1), ..., N(w_K)$ are collected
  - the geometric mean of the component frequencies is calculated $GM(f_1, ..., f_K) = (\prod_{k=1}^{K} N(f_k))^{\frac{1}{K}}$
  - compound word is split if $GM(f_1, ..., f_K) > N(f)$

The main difference between the two approaches is that the linguistic-based one leads to a larger number of split compounds because it does not depend on component frequencies, so even those compounds whose components have not been seen in the training will be split.

Examples of the splittings can be seen in Table 1. The first compound word "Arbeitnehmer" consists of two components, "Arbeit" and "Nehmer". Since the word "Nehmer" has not been seen in the training corpus, the geometric mean of component frequencies is equal to zero and therefore the word is not been split by the corpus-based method. The second compound word consists of three components, and each of them has been seen in the training corpus. However, the geometric mean of component frequencies is 17.9 whereas the frequency of the word itself is 51 which means that the word remains unsplit by the corpus-based method. Those values for the compound word "Treibhauseffekt" are also the reason for splitting the third word "Treibhauseffektgase" into two components instead of four.

Table 1. Examples of splitting German words

original word	splitted word	
	linguistic-based	corpus-based
Arbeitnehmer	Arbeit Nehmer	Arbeitnehmer
Treibhauseffekt	Treib Haus Effekt	Treibhauseffekt
Treibhauseffektgase	Treib Haus Effekt Gase	Treibhauseffekt Gase

2.2 Translation from English into German

For translation from English into German we propose three methods:

– splitting and merging German compounds
– POS-based joining of English words
– alignment-based joining of English words

Splitting and Merging German Compounds: German compound words in the training corpus are split using the corpus-based frequency method because it allows a straightforward and simple approach for merging components after the translation process. After training, translation is performed from English into the modified German language. The generated output is then postprocessed, i.e. the components are merged using the following method:

– a list of compounds and a list of components are extracted from the original German training corpus
– if the word in the generated output is in the component list
  • check if this word merged with the next word is in the compound list
  • if yes, merge two words

Joining English words: Another possible approach for treatment of the compound words in the target language is joining the corresponding words in the source language. Such transformation increases the English vocabulary size, but the word structure in the transformed English corpus becomes more similar to the German one.

– POS-based joining:
  English words which correspond to one German compound are usually two or more consecutive nouns. Therefore each sequence of English nouns is merged into one word.
– alignment-based joining:
  Distinct English words which are aligned to one German word are considered as potential components. All successive components are merged into one word.

Table 2. Examples of joining English words

original words	joined words	
	POS-based	alignment-based
energy certificate	energy_certificate	energy certificate
order of business	order of business	order_of business

As in the case of the German compound word splitting, the linguistic-based approach for joining English words (POS-based) leads to a larger number of English "compounds". An example can be seen in Table 2. The example shows two merged English nouns which have not been joined by alignment-based approach because in the baseline alignment they are not aligned to the same German word. The example also shows an aligment-based joining of a noun and a following preposition.

2.3 Improved Word Alignments

Knowlegde about splitting points of German compound words can also be used to enhance the word alignments. The alignments are trained using the modified German corpus with compound words split using the corpus-based frequency method described in Section 2.1. After the alignments are created, positions of the component words belonging to the same compound word are merged and

the training of translation models is done on the original German corpus. The advantage of this approach is that it can be applied to both translation directions without preprocessing of the input test text or postprocessing of the generated output.

3 Experiments

The experiments are performed on the European Parliament corpus described in [2]. It contains German and English parliamentary speeches. The corpus statistics can be seen in Table 3. The original corpus consists of about 700k sentences and 15M running words. In order to investigate effects of sparse training data, we have randomly extracted a small subset containing about 7k sentences and 144k running words (about 1% of the original corpus).

Table 3. Corpus statistics

		German	English
Train:	Sentences	751088	
	Running Words+Punctuation	15257678	16052330
	Vocabulary	205374	74708
	Singletons [%]	49.8	38.3
Dev:	Sentences	2000	
	Running Words+Punctuation	55147	58655
	Distinct Words	9213	6547
	OOVs [%]	0.8	0.2
Test:	Sentences	2000	
	Running Words+Punctuation	54260	57951
	Distinct Words	9048	6496
	OOVs [%]	0.7	0.2

As already pointed out, transformations were applied as a preprocessing step, then training and search were performed using the transformed data. In the case of improved alignments, the preprocessed corpus is used only for the alignment training, whereas the translation training is performed on the original corpus. The translation system we used is the phrase-based system described in [8]. Modifications of the training and search procedure were not necessary. In the case of the target language transformation, the inverse transformation step described in Section 2.2 was necessary after the translation.

The evaluation metrics used in our experiments are WER (Word Error Rate), PER (Position-independent word Error Rate) and BLEU (BiLingual Evaluation Understudy) [4].

4 Translation Results

4.1 Translation from German into English

Table 4 presents the results for translation from German into English. It can be seen that the treatment of German compound words leads to small but consistent improvements of all error measures for both sizes of the training corpus. The improvements obtained by the linguistic-based approach of compound splitting are similar to those of the corpus-based approach as well as to those obtained by improved word alignments.

Table 4. Translation results for German→English

German→English		dev			test		
		WER	PER	BLEU	WER	PER	BLEU
700k	baseline	63.4	48.6	20.5	63.6	48.6	20.9
	linguistic split	63.2	47.9	21.4	63.2	47.3	22.0
	corpus-based split	62.9	47.6	21.5	63.2	47.5	21.9
	improved alignment	63.1	48.4	21.1	63.3	48.3	21.5
7k	baseline	71.4	55.2	14.1	71.2	54.8	14.6
	linguistic split	71.5	54.5	15.0	71.1	53.7	15.6
	corpus-based split	71.3	54.5	15.0	71.0	53.7	15.4
	improved alignment	71.1	54.2	15.2	70.8	54.0	15.5

More details considering translation with the full corpus and corpus-based compound splitting are shown in Table 5.

Table 5. Detailed translation results for German→English

German→English			dev			test		
			WER	PER	BLEU	WER	PER	BLEU
700k	transformed	baseline	63.8	47.7	20.3	63.8	47.9	21.3
		split	63.2	46.5	21.4	63.4	46.6	22.5
	rest	baseline	63.0	49.3	21.0	63.4	49.2	20.8
		split	62.6	48.8	21.0	63.0	48.5	21.4

Development and test corpus were divided into two parts: one containing sentences with split compound words (which is about 45%)[1] and other which remained the same. Then these two sets were evaluated separately for each translation system. Results show that the compound splitting improves translation

[1] It should be noted that only about 2.5% of running words are affected by compound splitting, therefore significant changes in error measures cannot be expected.

quality for both sets, slightly more for the transformed set. This means that the new system allows better learning of models so that the translation quality has been improved both directly as well as indirectly.

Table 6. Translation examples for German→English without and with compound splitting

original German sentence:	...die artgerechte und umweltfreundliche **Produktionsmethode**...
transformed German sentence:	...die artgerechte und umweltfreundliche **Produktion Methode**...
generated English sentence:	
without splitting:	...the animal and environmentally friendly **production**...
with splitting:	...the animal and environmentally friendly **production methods**...
reference English sentence:	...the animal and environmentally friendly **production methods**...

From the translation example in Table 6 it can be seen that the system trained on the transformed corpus is better able to produce the correct English output.

4.2 Translation from English into German

The results for this translation direction are reported in Table 7.

Table 7. Translation results for English→German

English→German		dev			test		
		WER	PER	BLEU	WER	PER	BLEU
700k	baseline	68.6	56.4	19.8	68.5	56.2	19.8
	split+merge	68.4	55.9	20.4	68.3	55.5	20.4
	join-eng POS	68.5	56.1	20.1	68.2	55.5	20.6
	join-eng aligned	68.5	56.3	20.0	68.2	55.5	20.3
	improved alignment	68.2	55.9	20.2	67.7	55.2	20.6
7k	baseline	76.9	61.6	15.0	76.6	61.4	15.4
	split+merge	76.0	61.3	15.8	75.9	61.2	16.2
	join-eng POS	76.7	61.6	15.4	76.4	61.3	15.8
	join-eng aligned	76.8	61.8	15.2	76.4	61.4	15.8
	improved alignment	76.4	61.0	16.1	76.3	61.0	16.3

It can be seen that the treatment of German compounds is also helpful for this translation direction, namely when the German language is the target language.

For the full training corpus all four methods yield similar results, the splitting and merging method and the enhanced alignment yield slightly larger improvements. For the small training corpus, both methods for joining English words result to similar small improvements whereas the splitting and merging method and enhanced alignment have more impact.

Details for the translation with the full corpus and splitting-merging method can be seen in Table 8. Like for the other translation direction, the improvements are present for both evaluation sets, i.e. the translation quality has been improved both directly and indirectly.

Table 8. Detailed translation results for English→German

English→German			dev			test		
			WER	PER	BLEU	WER	PER	BLEU
700k	transformed	baseline	69.7	56.8	18.9	69.4	56.3	19.9
		split	69.2	56.0	19.9	69.3	55.5	20.5
	rest	baseline	67.5	56.4	20.4	67.4	56.4	19.6
		split	67.4	55.8	21.0	67.1	55.4	20.3

The translation example in Table 9 shows the advantage of the new system. Without compound treatment the system translated two English words belonging to one German compound into two German words. The output of the new system where German compounds have been split and merged is correct.

Table 9. Translation examples for English→German without and with compound splitting and merging

English sentence:	...the animal and environmentally friendly production methods...
generated German sentence:	
without splitting and merging:	...die artgerechte und umweltverträgliche **Produktion Methoden**...
with splitting and merging:	...die artgerechte und umweltfreundliche **Produktionsmethode**...
reference German sentence:	...die artgerechte und umweltfreundliche **Produktionsmethode**...

5 Conclusions

In this work we introduced several methods for dealing with German compound words in order to improve the translation quality of the German output. For translation from German into English we compared two approaches proposed in a previous work. We also proposed incorporating knowledge about German

compound words into the word alignments and tested it for both translation directions.

Our experimental results show that both linguistic-based and corpus-based compound splitting, as well as enhanced word alignment, yield similar improvements for translation from German into English.

It has been shown that these treatments of compound words also improve the quality of translation into German. For translation with a large tranining corpus, all proposed methods lead to similar improvements. For the small training corpus, splitting and merging German compounds and enhanced word alignment are slightly superior in comparison to the two other methods for joining English words.

In future work we plan to investigate possible treatments of compound words for other languages and language pairs (e.g. German-Spanish, Finnish, etc.). We also plan to investigate other methods for merging components in the generated output.

Acknowledgement

This work was partly supported by the TC-STAR project by the European Community (FP6-506738).

References

1. Koehn, P., Knight, K.: Empirical Methods for Compound Splitting. Proc. 10th Conf. of the European Chapter of the Association for Computational Linguistics (EACL). Budapest, Hungary (2003) 347–354
2. Koehn, P., Montz, C.: Shared task: statistical machine translation between European languages. Proc. ACL Workshop on Building and Using Parallel Texts. Ann Arbor, Michigan (2005) 119–124
3. Niessen, S., Ney, H.: Improving SMT quality with morpho-syntactic analysis. Proc. 18th Int. Conf. on Computational Linguistics (COLING). Saarbrücken, Germany (2000) 1081–1085
4. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: a method for automatic evaluation of machine translation. Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL). Philadelphia, PA (2002) 311–318
5. Popović, M., Ney, H.: Improving Word Alignment Quality using Morpho-syntactic Information. Proc. 20th Int. Conf. on Computational Linguistics (COLING). Geneva, Switzerland (2004) 310–314
6. Toutanova, K., Tolga Ilhan, H., Manning, C.: Extensions to HMM-based statistical word alignment models. Proc. Conf. on Empirical Methods for Natural Language Processing (EMNLP). Philadelphia, PA (2002) 87–94
7. Vilar, D., Matusov, E., Hasan, S., Zens, R., Ney, H.: Statistical Machine Translation of European Parliamentary Speeches. Proc. MT Summit X. Phuket, Thailand (2005) 259–266
8. Zens, R., Bender, O., Hasan, S., Khadivi, S., Matusov, E., Xu, J., Zhang, Y., Ney, H.: The RWTH Phrase-based Statistical Machine Translation System. Proc. Int. Workshop on Spoken Language Translation (IWSLT) Pittsburgh, PA (2005) 155–162