German Compounds in Factored Statistical Machine Translation

Sara Stymne

Department of Computer and Information Science
Linköping University
Sweden
sarst@ida.liu.se

Abstract. An empirical method for splitting German compounds is explored by varying it in a number of ways to investigate the consequences for factored statistical machine translation between English and German in both directions. Compound splitting is incorporated into translation in a preprocessing step, performed on training data and on German translation input. For translation into German, compounds are merged based on part-of-speech in a postprocessing step. Compound parts are marked, to separate them from ordinary words. Translation quality is improved in both translation directions and the number of untranslated words in the English output is reduced. Different versions of the splitting algorithm performs best in the two different translation directions.

1 Introduction

Compounding in German is productive and very common. Compounds are written without spaces or word boundaries. In statistical machine translation compounds lead to sparse data problems, increasing the number of unseen words. For translation into German it is a problem since several English words can be translated as distinct words rather than as a compound. To deal with these issues, compounds can be split into their component parts prior to training and translation, and for translation into German merged back together.

This study investigates how different compound splitting strategies influence factored phrase-based statistical machine translation (PBSMT). Translation between German and English is explored in both directions. An empirical method for compound splitting is used, which only requires a mono-lingual corpus and a part-of-speech (POS) tagger. Compound splitting and merging are performed as pre and postprocessing steps of the factored PBSMT system. Contrary to previous studies, parts of the split compounds are marked as such, to distinguish them from other words, since the semantics of compounds are not always compositional. Compounds are merged using a novel strategy based on part-of-speech.

Compound splitting is evaluated both on one-to-one correspondence with English and on translation quality. The main aims are to explore marked compound splitting and to find out which versions of an empirical compound splitting method give best results for translation of sentences in both directions between English and German in a factored PBSMT system.

2 Compounding

German compounds are formed by joining words without spaces or word boundaries. In addition, so called filler letters can occur between words, letters can be removed at the end of all but the last part of a compound, *umlaut* can be used, and there might be combinations of these. The term compound suffixes (*Kompositionssuffixen*) is used to describe these changes in [1], that also gives an overview of compound forms that occur in German noun compounds, based on a corpus study, summarized in Table 1.

Table 1. Compound suffixes in German.

Type	**Suffixes**	**Example**
None		Risikokapital (Risiko + Kapital) *risk capital*
Additions	*-s -n -en -nen* *-e -es -er -ien*	Arbeitsplatz (Arbeit + Platz) *Place of employment*
Truncations	*-e -en -n*	Südwesten (Süden + Westen) *south-west*
Combinations	*-us/-en -um/-en -um/-a* *-a/-en -on/-en -on/-a* *-e/-i*	Museenverwaltung (Museum + Verwaltung) *Museum management*
Umlaut	*umlaut + -er*	Völkermord (Volk + Mord) *genocide*

3 Related Work

German compounds in SMT have been addressed in a number of papers, (see e.g. [2–4]).

Translation into English is explored in [2], that use an empirical method where words are split in all possible parts, and for each part a check is performed against a monolingual corpus if it exists as an individual word. Additions of *-s* and *-es* are allowed to occur at all split points. A number of versions of the algorithm are tested in order to choose the correct splitting options, based on word frequencies, POS or bilingual alignment information. They find that an eager splitting method, choosing the splitting option with the highest number of splits, gives best translation results for PBSMT, despite having low precision and recall on one-to-one correspondence. A frequency-based ranking method based on the geometric mean of word frequencies gives similar results for PBSMT. Compound splitting also improves the translation quality of a word-based SMT

system. In this case using the geometric mean gives the best result, and the eager method gives worse result than no splitting at all. The systems are evaluated on NP/PPs, not on full sentences.

The same algorithm and in addition a rule-based method is used in [3]. In addition splitting is used only to improve word alignments. Both methods lead to improved translation quality. Both [2] and [3] integrate compound splitting by preprocessing training data and the text to be translated. No marking of compounds is used; the parts are treated as normal words.

In [3], compound splitting is also used for translation into German. They use the frequency-based version of the algorithm of [2] as in the other translation direction, and then merge compounds in a postprocessing step. The merging is based on two lists compiled from the German training corpus, a list of compounds and a list of compound components. If a word in the output is a compound component, they check if this word merged with the next is in the compound list, if it is, it is merged. A drawback of this method is that it only merges known compounds.

In addition [3] experiments with joining of English compounds based on POS or alignment data. All these methods lead to improved translation quality.

In [4], marking of split compounds is used in a factored PBSMT system with morphologically enriched POS-tags for German. A modified version of the splitting algorithm of [2] is used, which improved translation quality.

4 Processing of German Compound Words

German compounds are split in a preprocessing step and merged in a postprocessing step of translation.

4.1 Splitting Compounds

The splitting algorithm used in this study is the algorithm presented in [2], with a few modifications. Words are split in all possible places and a splitting option is chosen based on word frequencies from a monolingual corpus. The monolingual corpus is German Europarl text [5], with 1,467,291 sentences. It is POS-tagged and lemmatized using TreeTagger [6]. For the default algorithm the following changes from [2] have been made:

- The arithmetic mean of frequencies is used as default, rather than the geometric mean, in order to get more splits.
- Compound parts have to be of minimum three letters length.
- Words to be split are limited to content words: nouns, adjectives, adverbs and verbs. Proper names are, however, not split, since translating them in parts generally would give rise to errors.
- The last part of the compound must have the same POS as the full compound.
- The full list of compound suffixes in Table 1, except *umlaut*, is used

– In addition to surface form, lemmas are also used to calculate word frequencies. The reason for this is that compound parts often have the base form.
– Hyphenated words can only be split at hyphens.

The algorithm is varied by changing a number of parameters:

– The minimum length of words to be split and of compound parts is changed to 8 and 4 respectively.
– The scoring method is changed. In place of the arithmetic mean, the geometric mean of word frequencies is used or an eager method which choose the maximum number of parts. (These two methods are similar to the eager and frequency-based methods of [2])
– The number of parts per compound are restricted to maximum two and maximum two for all POS except nouns.
– All compound suffixes listed in Table 1, except *umlaut* are used, or only the 4 most common in the corpus study of [1], addition of *-s*, *-n*, *-en* or *-nen*.
– The restriction that the POS of the last part has to match the POS of the full compound is not used.

The splitting methods are summarized in Table 2. The methods differ in how many compounds they split, and in how many parts they split words, as is shown in Table 3, for the test text with a total of 55,580 words. The differences are large, with more than three times as many splits for the eager system as for the system with only common compound suffixes.

Table 2. Summary of the splitting options. The default method is shown with all settings, the other methods only show what differs from the default method.

Splitting	**Word length**	**Part length**	**Scoring**	**No. of parts**	**Suffixes**	**POS match**
default	6	3	arithm.	unlimited	all	yes
l8-s3	8					
l8-s4	8	4				
geom			geom.			
eager			eager			
nn2+				noun > 2		
max2				max 2		
common					common	
anypos						no

As pointed out in [2], parts of compounds do not always have the same meaning as when they stand alone. As an example they mention *Grundrechte* ('basic rights'), where the first part, *Grund*, usually translates as *foundation*, which is wrong in this compound.

To address this issue all compound parts but the last are marked with the symbol '#'. They are thus handled as separate words. Marking of parts also

means that they can keep their compound form, since they are not treated as normal words. If marking were not used it would be desirable to remove or add compound suffixes, as is done to a certain extent in [7].

Parts of split words also receive a special POS-tag, based on the POS of the last word of the compound, and the last part receives the same POS as the full word. (1) shows an example of a split word.

(1) Regierungskonferenz NN (*intergovernmental conference*) ⇒
Regierungs# NN-FL + Konferenz NN

Table 3. Number of compounds found in the test text for the different methods, showing both total, and number of parts per split compound.

Method	**Total**	**2**	**3**	**4**	**5+**
default	4862	3642	971	211	38
l8-s3	4624	3404	971	211	38
l8-s4	2985	2674	294	17	–
geom	3689	3063	534	74	18
eager	7729	4539	2050	788	352
nn2+	4693	3776	709	171	37
max2	4383	4383	–	–	–
common	2542	2303	228	11	–
anypos	6739	4691	1606	356	86

4.2 Merging Compounds

For translation into German a postprocessing step is performed where compounds are merged. Since a factored translation system is used, merging can be based on POS. If a word has a compound-POS, and the following word has a matching POS, they are merged. If the next POS is a conjunction, a hyphen is added to the word, allowing for coordinated compounds as in (2). Else the compound markup is simply removed. The POS-based algorithm has the advantage that it can merge unseen compounds and handle coordinated compounds.

(2) Wasser- und Bodenqualität
water and soil quality

4.3 Integration with Translation

The MT system used is a factored PBSMT system. In a factored system translation is not only based on surface form, but other features such as POS or lemma can be used in addition in different phases of translation (see [8]). The current

system uses POS as an output factor, and two sequence models, a 5-gram language model and a 7-gram POS-model, see Fig. 1. The Moses toolkit [9] is used for decoding and training, SRILM [10] for sequence models and Giza++ [11] for creating word alignments. Minimum error rate training [12] is used for tuning of feature weights. In addition German contracted prepositions and determiners are split in a preprocessing step, and for translation into German merged in connection with true casing by running a second Moses instance.

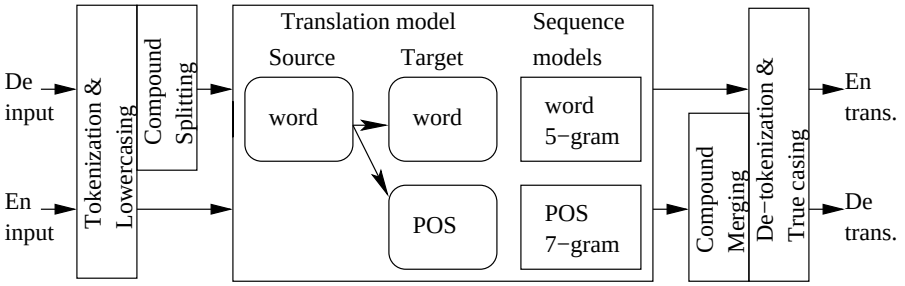

Fig. 1. Architecture of the factored system

All corpora are European Parliament texts [5]. The size of the corpora are 439,513 sentences for training, 600 sentences for tuning and 2000 sentences for testing[1].

Compound splitting is integrated in a preprocessing step for training and for translation from German. Compounds are split using the different versions of the splitting algorithm described in Sec. 4.1. For translation into German, compounds are merged in a postprocessing step.

5 Evaluation

The nine splitting methods described in Sec. 4.1 are compared to each other and to a baseline without splitting (*raw*) for one-to-one correspondence with English and for translation quality.

5.1 One-to-One Correspondence

To measure one-to-one correspondence I followed the evaluation method described in [2]. One-to-one correspondence occurs when the words in a German compound are translated as separate content words in English. In addition there can be inserted function words. As an example *Medienfreiheit* is in one-to-one

[1] The test set is *test2007* from the ACL 2008 Workshop on Statistical Machine Translation, http://www.statmt.org/wmt08/shared-task.html

correspondence with *freedom of the media*, since the two German parts *Medien* and *Freiheit* corresponds to two separate words, *media* and *freedom*. The lack of correspondence of the two function words, *of the*, is not considered.

A gold standard was created by manually annotating the first 5000 words of the test text for one-to-one correspondence with the English reference text. Out of the 5000 words, 174 were compounds in one-to-one correspondence with English.

The results of the one-to-one evaluation is shown in Table 4. The same categories and metrics as in [2] are used:

correct split: words that were correctly split
correct not: words that should not be split and were not
wrong not: words that should be split but were not
wrong faulty: words that were split but in an incorrect way
wrong split: words that should not be split but were
precision: (correct split) / (correct split + wrong faulty + wrong split)
recall: (correct split) / (correct split + wrong faulty + wrong not)
accuracy: (correct) / (correct + wrong)

Table 4. One-to-one correspondence of split compounds compared to a manually annotated gold standard for the different splitting methods.

Method	**Correct**		**Wrong**			**Metrics**		
	split	**not**	**not**	**faulty**	**split**	**prec.**	**recall**	**acc.**
raw	0	5000	174	0	0	–	0.0%	**96.6%**
default	99	4504	22	52	323	20.9%	57.2%	92.1%
l8-s3	99	4530	22	52	297	22.1%	57.2%	92.6%
l8-s4	120	4692	36	17	135	**44.1%**	69.3%	96.2%
geom	109	4614	33	31	213	30.9%	63.0%	94.5%
eager	43	4243	18	112	584	5.8%	24.9%	85.7%
nn2+	99	4521	24	50	306	21.8%	57.2%	92.4%
max2	133	4546	29	11	281	31.3%	**76.9%**	93.6%
common	99	4714	58	16	113	43.4%	57.2%	**96.3%**
anypos	99	4310	10	64	517	14.6%	57.2%	88.2%

The splitting options have their strengths on different metrics, with three different methods having the best results for the three metrics used.

Compared to the default method it can be seen that both imposing length restrictions and using the geometric mean increases the results on all three metrics. Limiting the number of parts makes a minor difference when nouns are excluded, but gives the highest recall when used for all POS. Using only common compound suffixes gives higher precision, whereas not using the POS restriction on the last word gives lower precision. No splitting actually gives the highest accuracy.

The largest error category is wrong splits. The splits in this category are reasonable, in the sense that all parts are meaningful German words, even if they are not in one-to-one correspondence with English. As an example, of the 323 wrong splits for the default system, 234 (72,5%) are reasonable. The erroneous splits often have parts that are common words such as *ich* ('I') and *ist* ('is').

Compared to [2], the two similar systems, eager and geom, have lower results on all metrics. This might in part be due to other changes made to the algorithm, such as allowing more compound suffixes, but can also be because these algorithms make more mistakes on full sentences than on NP/PPs.

5.2 Translation Quality

Translation quality is measured against one reference translation, using three metrics, BLEU [13], NIST [14] and METEOR [15].[2]

German ⇒ English

The results for translation from German into English can be seen in Table 5[3]. All systems with compound splitting get higher NIST and METEOR scores than the raw system, but only the geom system has a higher BLEU score than the raw system. The geom system, which is similar to the frequency-based system, that performed well in [2], has the highest score for all metrics. The eager system, however, performs poorly. This is probably because it makes more mistakes when used on full sentences than on only NP/PPs.

Table 5. Translation results for German ⇒ English

Method	BLEU	NIST	METEOR
raw	26.29	6.888	52.27
default	26.12	6.915	52.62
l8-s3	26.13	6.920	52.61
l8-s4	26.20	6.935	52.61
geom	**26.35**	**6.945**	**52.79**
eager	25.88	6.898	52.45
nn2+	26.10	6.923	52.61
max2	26.23	6.934	52.67
common	26.22	6.944	52.54
anypos	26.12	6.920	52.59

[2] The evaluation is case-sensitive. %BLEU and %METEOR notation is used. METEOR is used with the "exact" and "porter stem" modules, the WordNet-based modules for English are not used.

[3] As [3] note, only a small percentage of words are affected by compound splitting so significant changes in error measures can not be expected.

Two other systems that perform reasonably well are the systems with common compound suffixes and maximum 2 splits. Of these the common system has high precision and accuracy on the one-to-one evaluation and the max2 has high recall. Limiting the length of both words to be split and compound parts gives rise to small improvements in BLEU and NIST.

One improvement that can be seen in the systems with split compounds is that the number of untranslated words is reduced by more than half. The raw system has 733 untranslated words (1.25%), compared to 360 words (0.61%) in the geom system. Of the untranslated words in the geom system 45 (12.5%) are marked compound parts, which could possibly have been translated if marking were not used. Among the other untranslated words there are many proper names and unsplit compounds. The translation example in Table 6 shows an example of a sentence where the systems that split compounds, exemplified by geom, manages to translate a compound that is untranslated by the raw system.

Table 6. Sample translation for German ⇒ English with and without compound splitting

Sentence type	**Example**
De original	... der Koordinierung der Außen- und **Sicherheitspolitiken**...
De preprocessed	... der koordinierung der außen- und **sicherheits# politiken**...
En with splitting	... the coordination of the foreign and **security policies**...
En without splitting	... the coordination of the foreign and **sicherheitspolitiken**...
En reference	... to coordinate the common foreign and **security policies**...

English ⇒ German

The result for translation from German into English can be seen in Table 7. In this direction the systems with splitting had higher scores than the raw system for all metrics and systems, except the eager system for BLEU. The eager system had the worst performance of the systems with splitting in this translation direction as well.

The best scoring systems in this direction are not the same as in the opposite direction. The default system and nn2+ had the highest scores. These systems have lower precision, just over 20%, on the one-to-one evaluation, than the systems that performed best in the opposite direction.

In this direction, imposing length limits on words to be split and compound parts led to worse translation results, as opposed to the other direction where it improved the results.

Table 7. Translation results for English ⇒ German

Method	BLEU	NIST	METEOR
raw	19.31	5.727	26.53
default	**19.73**	**5.854**	27.05
l8-s3	19.63	5.833	27.02
l8-s4	19.56	5.821	26.96
geom	19.64	5.818	26.95
eager	19.16	5.788	26.75
nn2+	19.71	5.850	**27.07**
max2	19.66	5.837	26.98
common	19.67	5.824	27.03
anypos	19.62	5.853	27.01

An example where the systems that split compounds handle compounds better can be seen in Table 8, exemplified by the default system. The default system manages to produce the desired compound, whereas the raw system produces two nouns instead.

Table 8. Sample translation for German ⇒ English with and without compound splitting

Sentence type	Example
En original	... of the **national states** than to represent genuine progress...
De with splitting	... der|art **national#|nn-fl staaten|nn** als|kokom echte|adja fortschritte|nn zu|ptkzu machen|vvfin...
De with splitting, postprocessed	... der **Nationalstaaten** als echte Fortschritte zu vertreten...
De without splitting	... von den **Nationalen Staaten** als echte Fortschritte zu machen...
De reference	... der Nationalstaaten zu bekräftigen, als dass sie einen wirklichen Fortschritt darstellt...

5.3 Discussion

The methods that improved translation quality most were different in the two translation directions. A method using the geometric mean of word frequencies performs best for translation into English, and limiting the number of splits to two and only using common compound suffixes also performs well. Methods using the arithmetic mean of word frequencies, and limiting the number of splits to two for all words but nouns worked best for translation into German. Limiting the number of compound suffixes gives good results in both directions.

Generally systems with more total splits perform better for translation into German, and systems with fewer splits perform better for translation into English.

One-to-one correspondence does not seem to be a good indicator for judging if a splitting method will improve PBSMT. In part this could be explained by the fact that the PBSMT system aligns word sequences, and thus can rejoin split words in the translation model. Another reason can be that a larger number of splits increases the chance of splitting unseen compounds into known parts at translation time.

Since compounds only make up a small proportion of all words the differences found between systems were small in many cases. Human analysis of translation output will be needed to shed further light on these small improvements. A small qualitative study of compound translation for a system using a similar splitting method indicates that translation of compounds is improved by splitting compounds [4].

6 Conclusion

A number of versions of an empirical compound splitting method have been explored for translation between German and English in both directions. Incorporating them into a factored translation system and marking compounds did give a small improvement of translation quality. Particularly the number of untranslated words are reduced by approximately half for translation into English. However, marking does lead to a small number of untranslated compound parts.

As in previous work, methods with high scores on metrics for one-to-one correspondence with English did not give the best translation results for German to English. This study shows that the same holds for translation in the opposite direction.

This study has also indicated that to achieve good translation results splitting should not necessarily be performed using the same method for translation in different directions.

Some of the methods that worked well have not yet been tried in combination, which would be interesting in future work. The methods can also be expected to work well for other compounding languages, such as Swedish or Italian.

References

1. Langer, S.: Zur Morphologie und Semantik von Nominalkomposita. In: Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache. (1998) 83–97
2. Koehn, P., Knight, K.: Empirical methods for compound splitting. In: Proceedings of the tenth conference of EACL, Budapest, Hungary (2003) 187–193
3. Popović, M., Stein, D., Ney, H.: Statistical machine translation of German compound words. In: Proceedings of FinTAL - 5th International Conference on Natural Language Processing, Turku, Finland (2006) 616–624

4. Stymne, S., Holmqvist, M., Ahrenberg, L.: Effects of morphological analysis in translation between German and English. In: Proceedings of the Third ACL Workshop on Statistical Machine Translation, Columbus, Ohio (2008)
5. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit X, Phuket, Thailand (2005) 79–86
6. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK (1994) 44–49
7. Holmqvist, M., Stymne, S., Ahrenberg, L.: Getting to know Moses: Initial experiments on German-English factored translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic (2007) 181–184
8. Koehn, P., Hoang, H.: Factored translation models. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic (2007) 868–876
9. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL, demonstration session, Prague, Czech Republic (2007) 177–180
10. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Denver, Colorado (2002) 901–904
11. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
12. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting of ACL, Sapporo, Japan (2003) 160–167
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, Pennsylvania (2002) 311–318
14. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, San Diego, California (2002) 138–145
15. Lavie, A., Agarwal, A.: METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic (2007) 228–231