Processing of Swedish Compounds for Phrase-Based Statistical Machine Translation

Sara Stymne and Maria Holmqvist

Department of Computer and Information Science
Linköping University, Sweden
{sarst,marho}@ida.liu.se

Abstract. We investigated the effects of processing Swedish compounds for phrase-based SMT between Swedish and English. Compounds were split in a pre-processing step using an unsupervised empirical method. After translation into Swedish, compounds were merged, using a novel merging algorithm. We investigated two ways of handling compound parts, by marking them as compound parts or by normalizing them to a canonical form. We found that compound splitting did improve translation into Swedish, according to automatic metrics. For translation into English the results were not consistent across automatic metrics. However, error analysis of compound translation showed a small improvement in the systems that used splitting. The number of untranslated words in the English output was reduced by 50%.

1 Introduction

In many languages, including the Germanic languages, compounding is very common, and compounds are written without spaces or other word boundaries. This is problematic for many NLP applications. For phrase-based statistical machine translation (PBSMT) it leads to problems due to data sparseness, with a large number of out-of-vocabulary compounds.

This problem has been studied in several papers for translation between German and English. Koehn and Knight [1] suggested an empirical algorithm for splitting compounds, which was successfully applied to German-to-English translation. The same method was used by Popović et al. [2] for translation in both directions between English and German. In addition, they merged compounds in a postprocessing step for translation into German. Stymne [3] tried a number of variations of the algorithm for translation in both translation directions. In both studies translation quality was improved.

Compound parts are usually treated as ordinary words in the training data, e.g. in [1, 2]. In [3, 4], however, compound parts were marked with a symbol, to separate them from normal words, resulting in improved translation quality compared to an unsplit baseline.

Virpioja et al.[5] used an unsupervised algorithm for morphological splitting and merging, where both compounds and other words were split into stems and affixes, for translation between Swedish and other Scandinavian languages.

Table 1. Compound forms in Swedish

Type	**Suffixes**	**Example**
None		riskkapital (risk + kapital) *risk capital*
Additions	*-s -t*	frihetslängtan (frihet + längtan) *longing for peace*
Truncations	*-e -a*	pojkvän (pojke + vän) *boyfriend*
Combinations	*-a/-s -a/-e -a/-u -a/-o -e/-a -e/-s -el/-la -la/-el -ra/-er*	arbetsgrupp (arbete + grupp) *working group*

All but the last part of a word were marked with a symbol, which was used in the merging step. They did not get improved translations when measured automatically, but saw other advantages such as a reduction of out-of-vocabulary words.

There have been many other suggestions of how to split compounds, but they are often only evaluated against a gold standard, not on a translation task. Alfonseca et al. [6] suggested a language independent supervised learning method, which needs a corpus of annotated compounds. They also showed that the training corpus can be of another language than the corpus to be split. For Swedish, Sjöbergh and Kann [7] suggested several ways of choosing the correct splitting points of compounds that were split using a method based on word lists.

The corpus-based language independent compound splitting method suggested in [1] was shown to be useful for PBSMT from and into German. In this work we investigate if a similar empirical method is useful for translation from and into Swedish. In addition we investigate the effects of marking compound parts [3], compared to the more commonly used strategy where no marking of compound parts is used. We also present a novel POS- and corpus-based merging algorithm for compounds.

2 Swedish Compounds

Compounds in Swedish are normally formed by joining words, without any spaces or other word boundaries. Compound parts can have special compound forms, created by addition of letters, truncation of letters or combinations of these. An overview of compound forms can be seen in Table 1, compiled from two standard works on Swedish morphology [8, 9].

In addition to the forms in Table 1, the spelling of compounds is changed in cases where adding two words would result in three identical consecutive consonants. In such a case one of the three consonants is removed, leaving a double consonant. An example of this can be seen in (1). This can lead to ambiguities, as in (1), which usually can be easily disambiguated semantically by a human, but which can lead to problems for automatic splitting methods.

(1) stopplikt – stopp + plikt / stop + plikt
obligation to stop – stop + duty / stoup + duty

2.1 Compound Splitting

To handle compounds in PBSMT we split them into their parts. We use a slightly modified version of an empirical splitting method based on [1, 3].

For each word all possible splits of the word are tried, and a split is considered if all its parts are found as words in a monolingual corpus. If there are several candidates the splitting option with the highest arithmetic mean of the frequencies of its parts is chosen, which can be the original unsplit word if it is common. We also use part-of-speech information, from the Swedish Granska tagger [10]. We retokenize the tagger output to split word groups that are tokenized as one item by the tagger, such as time expressions and coordinated compounds.

We split nouns, verbs, adjectives and adverbs, which are the parts-of-speech that form compounds in Swedish. Proper names are excluded, since they generally are not translated in parts, as the Swedish surname *Sjögren*, the parts of which mean *lake* and *branch*. The same parts-of-speech plus proper names and numerals are used for frequency calculations from the monolingual corpus.

We also impose a restriction that the last part of the compound must have the same part-of-speech as the full compound. In addition to surface forms, base forms, obtained from the tagger, are also used for frequency calculations, since compound parts tend to have base form. We also impose limits on length, for a word to be split it must have at least six characters and each part must have at least three characters. Additions of *-s* and truncations of *-e* and *-a* are allowed at all split points[1]. We also handle cases with consecutive consonants, by allowing the addition of an extra consonant at splitting points between two identical consonants.

We use two schemes to handle compound parts, *marked* and *unmarked*. In the marked scheme compound parts keep the form they have in the compound, except in the three consonant case, where a consonant is added. In addition we add the symbol '#' to all but the last part, to separate compound parts from other words, since compounds are not always compositional in meaning. In the unmarked scheme we normalize compound parts to a canonical form based on the suffixes in Table 1, and no marking is added. The canonical form will coincide with a word form that occurs independently in the corpus, or the base form of such a word. We give the last part of the compound the same POS-tag as the full word, whereas the other parts get a special tag, based on the original tag. An example of the splitting schemes is shown in (2).

(2) Compound förvaltningssystem NN
administrative system
Unmarked förvaltning NN-FL + system NN
Marked förvaltnings# NN-FL + system NN

[1] Using more variants of compound forms was not successful in a small pilot study.

2.2 Compound Merging

For translation into Swedish it is necessary to merge compounds that are translated in parts. For marked compounds we use the POS-based algorithm suggested in [4]. If a word has the special part-of-speech used for compound parts, it is merged with the following word if it has a matching part-of-speech, which can either be another compound part, or the final part of a compound. In addition we handle coordinated compounds, as in (3), by adding a hyphen to a word when the next word is a conjunction. In cases where the merged words would have three identical consecutive consonants, see (1), we remove a consonant. No other processing is needed in this setting, since compound forms are kept, except removing compound markup.

(3) kunskaps- och informationssamhälle
knowledge and information society

For unmarked compounds a more elaborate strategy is needed to handle the normalized compound forms. In addition to part-of-speech, we use frequency lists of all words from the training corpus, and of compound parts with all possible compound forms found during splitting. To find the correct form of a word we first try all combinations of forms of each compound part and check if the result is a word that is known from the corpus. If any known word is found we choose the most frequent one. Else, we add the parts from left to right choosing the most frequent possible combination at each merging point, and if no known combination exists, the most frequent compound form for each part.

To investigate the potential of the merging method for unmarked compounds we applied it to the split test text (see section 3.1). The splitting algorithm found 2505 compounds, of which 227 are out-of-vocabulary with respect to the training corpus. The merging method correctly merged all but 91 compounds, showing that it finds all known compounds and have a reasonable success (60%) on unknown compounds. Most incorrect compounds can easily be understood by a human, even if the form is wrong. The most common error is a left out addition of *-s*.

3 System Description

The translation system we use is a factored phrase-based SMT system, with part-of-speech as an additional output factor. We use TreeTagger [11] to tag the English texts and Granska tagger [10] to tag the Swedish texts. We use two sequence models, produced by SRILM [12], a 5-gram language model on surface form and a 7-gram model on part-of-speech. For training and decoding we use the Moses toolkit [13]. We tune feature weights using minimum error rate training [14], that optimizes the Neva metric [15].

A pre-processing step is performed on the Swedish side where compounds are split. Thus we train the system on English and modified Swedish. Compounds are merged after translation into Swedish and during tuning.

Table 2. Number of tokens and types for the training corpus

System	Tokens	Types
Baseline	13603062	182000
Swedish Unmarked	14401784	100492
Marked	14401784	107047
English	15043321	67044

We train three systems for this study, a baseline system without splitting, and two systems with splitting, *marked*, where compound parts are marked and *unmarked*, with unmarked parts normalized to canonical form.

3.1 Corpus

The translation system is trained and tested using the Europarl corpus [16]. The training part contains 701157 sentences, where sentences longer than 40 words have been filtered out. The number of tokens and types in the training corpus is shown in Table 2. The Swedish baseline text contain 2.7 times as many types as the English side. Splitting Swedish compounds reduces the vocabulary size by up to 45%. The development and test corpora are taken evenly from the designated test portion of the fourth quarter of 2000. The test set has 2000 sentences and the development set has 500 sentences.

4 Evaluation of Compound Splitting

We use two manually created gold standards to evaluate compound splitting. The gold standard corpus consists of the first 5395 words (245 sentences) from the test set.

4.1 Gold Standards

For the first gold standard all compounds[2] in the gold standard corpus were annotated. This standard was prepared by two human judges who are native speakers of Swedish.

To investigate the difficulty of the task we calculated agreement as suggested in [6], as the percentage of agreement in classification as compounds or non-compounds (CCA), the Kappa score [17] obtained from CCA and the percentage of words for which the suggested decomposition was identical (DA). Since we evaluate on running text, which has a very large percentage of non-compound words, the agreement could be expected to be high. Therefore we also measured agreement on only those words that are 12 characters or longer, to have a more

[2] A word is considered to be a compound if it has several parts which all are semantically meaningful with respect to the full compound and can be used as stand alone words in some form.

Table 3. Inter-judge agreement scores for compound classification

Type	CCA	Kappa	DA
Full test	98.2%	0.96	98.0%
Long words	91.1%	0.82	97.7%

Table 4. Results of compound splitting

Test set		Prec	Rec	Acc
All compounds	full	56.4%	53.0%	95.8%
	long	76.6%	51.3%	76.8%
One-to-one	full	31.9%	66.4%	96.1%
	long	55.5%	65.7%	81.4%

even distribution of compounds and non-compounds. As shown in Table 3, the agreement is high for all metrics and both samples.

For the final evaluation, the two judges agreed on a common judgement for the words where they disagreed. The final test text has 288 compounds out of 5395 words. We also used the test set with words that are 12 character or longer, that contains 626 words, of which 231 are compounds.

The second gold standard consists of Swedish compounds whose parts are in one-to-one agreement with separate words in the English translation, allowing insertion of function words [1]. Since this task is more straightforward than for all compounds, which has a high inter-judge agreement, only one judge created this gold standard. It contains 126 compounds out of 5395 words. Again we also use the test set with long words, which contains 626 words, of which 117 are compounds in one-to-one agreement with English.

4.2 Results

The evaluation uses the three metrics precision, recall and accuracy, as defined by [1]. Table 4 shows the result of the evaluation of the compound splitting. Precision is higher for all compounds and recall is higher on the one-to-one test set, which is quite natural, considering that the number of compounds is much smaller in the one-to-one test set. On the test set with only long words precision shows a big improvement, recall a small drop, and accuracy a large drop over the full test set.

Comparing the splitting accuracy to other studies, it is worse on linguistic evaluation than both the supervised method of [6] and the word list based method of [7], where, however, only accuracy on a corpus of only compounds is measured. It does perform better than some simpler versions of the algorithm in [6], e.g. their reimplementation of [1], that are only tested on German. We also have worse precision and recall on 1-to-1 evaluation than the similar frequency-based method used for German [1], who evaluated only on NP/PPs. However, better results on these metrics did not necessarily give better translation quality for PBSMT [1, 3], probably because phrases with compounds that were erroneously split were linked together in the training phase.

5 Evaluation of Translation

Translation was evaluated both by automatic measures and by human error analysis, which focus on out-of-vocabulary words and translation of compounds.

Table 5. Results for translation from Swedish to English

System	Meteor	Bleu	Neva	NIST
Baseline	55.47	29.97	34.08	7.3127
Unmarked	55.82	29.89	34.08	7.3470
Marked	55.78	29.85	34.05	7.2933

Table 6. Results for translation from English to Swedish

System	Meteor	Bleu	Neva	NIST
Baseline	57.86	21.63	26.53	6.1085
Unmarked	58.43	22.12	26.99	6.1430
Marked	58.31	21.92	26.81	6.2025

5.1 Automatic Metrics

We use four automatic metrics, Bleu [18], NIST [19], Neva [15] and Meteor[3] [20]. Case sensitive versions of the metrics are used.

The result for translation from Swedish can be seen in Table 5. The unmarked system is slightly better than the baseline on Meteor and NIST, but worse on Bleu. The marked system is worse than the unmarked on all metrics, and only better than the baseline on Meteor.

Table 6 shows the results for translation into Swedish. In this direction the differences between the scores are bigger and both split systems beat the baseline on all metrics. The unmarked system is better than the marked system on all metrics except NIST.

5.2 Out-of-Vocabulary Words

To investigate the effects of compound splitting on translation from Swedish we analysed the out-of-vocabulary (OOV) words in the systems. These words are left untranslated in the system output. The total number of out-of-vocabulary words are reduced by about 50% in the split systems, compared to the baseline. A manual analysis showed that this decrease was to the largest part due to a higher proportion of translated compounds, see Table 7. The system with marked compounds has a slightly higher number of OOV:s, mainly due to the fact that 16 marked compound parts are left untranslated. Of the remainder of the OOV:s in the unmarked system, 55 are numerals, 27 are proper names, 7 foreign words, and 82 miscellaneous unseen words.

5.3 Compound Translation from Swedish

To investigate compound translation from Swedish we manually evaluated the translation of the first 100 compounds in the test text, with a clear translation in the English reference text. We then classified the translations in the test text with respect to the reference text. The result can be seen in Table 8. As expected the number of OOV:s is reduced in the systems with splitting. There is a small increase in the number of compounds that are identical to or a good alternative

[3] For English a version of Meteor optimized on human judgements is used, for Swedish the original Meteor weights are used. For both languages the "exact" and "porter stem" modules are used.

Table 7. Unique out-of-vocabulary words and compounds, for translation into English

System	Comps/OOV:s
Baseline	331/520
Unmarked	87/258
Marked	86/269

Table 8. Analysis of translation of 100 compounds from Swedish

	Baseline	Unmarked	Marked
Identical	58	59	60
Alternative	19	21	21
Understandable	7	13	11
Partly transl.	2	1	1
Missing	1	2	3
Wrong	3	2	2
OOV	10	2	2

Table 9. Analysis of translation of 100 compounds into Swedish

	Baseline	Unmarked	Marked
Identical comp.	48	53	57
Alt. compound	14	9	10
Alt. word	16	16	12
Alt. word group	9	8	9
Split compound	7	5	3
Partly transl.	4	7	4
Missing	0	0	2
OOV	2	2	3

to the reference translation in the systems with splitting. The largest increase, however, is in translations that convey the meaning but is somewhat ill-formed, the understandable category.

These results are not as promising as in similar evaluations for German [4], which used similar compound splitting strategies.

5.4 Compound Translation into Swedish

We performed a similar evaluation in the opposite translation direction, using the same sample of 100 compounds from the reference translation. In this direction the categories were changed slightly. For the alternative translations, we also distinguished between translation that were compounds, single words or word groups. There is also a category for word groups that were translated as separate words, but should have been compounded.

The result of this evaluation can be seen in Table 9. There are more translations that are identical to the reference in the two systems with splitting, but the total number of identical and alternative translations are approximately the same in the three systems. The number of split compounds is higher in the baseline system. The unmarked system produces more split compounds and partial translations than the marked system. This can be seen as an indication of marking having an effect, which, however, is not seen in the automatic evaluation.

No merging errors were found in this sample for the marked system. In the unmarked system the merging algorithm performed correctly for 60 of the 62

merged compounds. Of the two errors, the first, (4a), is a missing addition of an *-s* and the second error, (4b), is not covered by the algorithm since it should have been a combination of *-e/-s*, and combinations are not handled. The presence of such a word in this small sample indicates that it is worth investigating allowing more compound forms in the future.

(4) a. *medlemländer
medlemsländer – medlem + länder
member states – member + countries

b. *samhällepolitiska
samhällspolitiska – samhälle + politiska
socio-political – society + political

6 Conclusions

In this study we have investigated the effects of splitting and merging Swedish compounds for PBSMT between Swedish and English. An unsupervised empirical compound splitting method is used. Even though the splitting method does not have a particularly high precision and recall compared to any of the two gold standards created, when incorporated into translation, it still improves automatic scores for translation into Swedish. For translation from Swedish the automatic metrics are inconsistent. In both directions, the error analysis shows a small improvement of compound translation.

A big improvement for translation into English is that the number of out-of-vocabulary words, that leads to untranslated words in the translation output, is reduced by approximately half. There are, however, still some untranslated compounds left, which indicates that it might be useful to apply a more advanced and resource intensive splitting strategy (e.g. [6, 7]) for PBSMT.

Measured by automatic metrics the system that uses canonical form of compound parts is generally better than the system that uses marked compound parts. In the error analysis, the difference between the two versions are smaller, with the marked system being slightly better in some cases. A drawback of the marked system is that it has a small number of untranslated marked compound parts.

The two suggested merging algorithms work well, and generally produce valid Swedish compounds. In a few cases the merging method for unmarked compounds produces incorrect compound forms of parts. The resulting words are usually understandable for a human, and are better translation alternatives than untranslated words.

Compared to [5], where both compounds and other words were split into stems and affixes, we find our results more promising. In contrast to their results, we do see some improvements using automatic metrics. The results are not directly comparable since different language pairs are used, but a similarity is the large reduction of untranslated words in the output. Encouraged by these results, our aim is to further explore compound processing for PBSMT, since we believe it will lead to improved translation quality.

References

1. Koehn, P., Knight, K.: Empirical methods for compound splitting. In: Proc. of EACL-03, Budapest, Hungary (2003) 187–193
2. Popović, M., Stein, D., Ney, H.: Statistical machine translation of German compound words. In: Proc. of FinTAL, Turku, Finland (2006) 616–624
3. Stymne, S.: German compounds in factored statistical machine translation. In: Proc. of GoTAL, Gothenburg, Sweden (2008) 464–475
4. Stymne, S., Holmqvist, M., Ahrenberg, L.: Effects of morphological analysis in translation between German and English. In: Proc. of the Third Workshop on Statistical Machine Translation, Columbus, Ohio (2008) 135–138
5. Virpioja, S., J.Väyrynen, J., Creutz, M., Sadeniemi, M.: Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In: Proc. of MT Summit XI, Copenhagen, Denmark (2007) 491–498
6. Alfonseca, E., Bilac, S., Pharies, S.: Decompounding query keywords from compounding languages. In: Proc. of ACL-08: HLT, Short Papers, Columbus, Ohio (2008) 253–256
7. Sjöbergh, J., Kann, V.: Finding the correct interpretation of Swedish compounds, a statistical approach. In: Proc. of LREC, Lisbon, Portugal (2004) 899–902
8. Thorell, O.: Svensk ordbildningslära. Esselte Studium, Stockholm, Sweden (1981)
9. Hellberg, S.: The Morphology of Present-Day Swedish. Number 13 in Data linguistica. Almqvist & Wiksell, Stockholm, Sweden (1978)
10. Carlberger, J., Kann, V.: Implementing an efficient part-of-speech tagger. Software Practice and Experience **29** (1999) 815–832
11. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proc. Intl. Conf. on New Methods in Language Processing, Manchester, UK (1994) 44–49
12. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proc. Intl. Conf. on Spoken Language Processing, Denver, Colorado (2002) 901–904
13. 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: Proc. of ACL-07, demo session, Prague, Czech Republic (2007) 177–180
14. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proc. of ACL-03, Sapporo, Japan (2003) 160–167
15. Forsbom, E.: Training a super model look-alike: featuring edit distance, n-gram occurrence, and one reference translation. In: Proc. of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation, New Orleans, Louisiana (2003) 29–36
16. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Proc. of MT Summit X. (2005) 79–86
17. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics **22**(2) (1996) 249–254
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proc. of ACL-02, Philadelphia, Pennsylvania (2002) 311–318
19. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proc. of the Second Int. Conf. on Human Language Technology, San Diego, California (2002) 228–231
20. Lavie, A., Agarwal, A.: METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proc. of the Third Workshop on Statistical Machine Translation, Prague, Czech Republic (2007) 228–231