Paraphrasing Swedish Compound Nouns in Machine Translation

Edvin Ullman and **Joakim Nivre**
Department of Linguistics and Philology, Uppsala University
edvinu@stp.lingfil.uu.se joakim.nivre@lingfil.uu.se

Abstract

This paper examines the effect of paraphrasing noun-noun compounds in statistical machine translation from Swedish to English. The paraphrases are meant to elicit the underlying relationship that holds between the compounding nouns, with the use of prepositional and verb phrases. Though some types of noun-noun compounds are too lexicalized, or have some other qualities that make them unsuitable for paraphrasing, a set of roughly two hundred noun-noun compounds are identified, split and paraphrased to be used in experiments on statistical machine translation. The results indicate a slight improvement in translation of the paraphrased compound nouns, with a minor loss in overall BLEU score.

1 Introduction

Swedish is a highly productive language, new words can be constructed fairly easily by concatenating one word with another. This is done across word classes, although, as can be expected, predominantly with content words. Due to this high productivity, an exhaustive dictionary of noun compounds in Swedish does not, and can not exist. Instead, in this project, noun compounds are extracted from the Swedish Europarl corpus (Koehn, 2005) and a subset of Swedish Wikipedia,[1] using a slight modification of the splitting method described in Stymne and Holmqvist (2008), based on previous work by Koehn and Knight (2003).

The assumption that paraphrases of noun compounds can help in machine translation is supported in Nakov and Hearst (2013). Although this study was conducted with English compound nouns, a similar methodology is applied to the Swedish data. The split compound nouns are paraphrased using prepositional and verb phrases, relying on native speaker intuition for the quality and correctness of the paraphrases. A corpus is then paraphrased using the generated paraphrases and used to train a statistical machine translation system to test whether or not an improvement of quality can be observed in relation to a baseline system trained on the unmodified corpus. The results show a minor improvement in translation quality for the paraphrased compounds with a minor loss in overall BLEU score.

2 Background

Previous studies on the semantics of compound nouns have, at least for the English language, in general focused on finding abstract categories to distinguish different compound nouns from each other. Although different in form, the main idea is that a finite set of relations hold between the constituents of all compound nouns. Experiments have been done to analyse such categories in Girju et al. (2005), and applied studies on paraphrasing compound nouns with some form of predicative representation of these abstract categories were performed in Nakov and Hearst (2013).

Studies on Swedish compound nouns have had a slightly different angle. As Swedish noun compounding is done in a slightly different manner than in English, two nouns can be adjoined to form a third, two focal points in previous studies have been detecting compound nouns (Sjöbergh and Kann, 2004) and splitting compound nouns (Stymne and Holmqvist, 2008; Stymne, 2009).

Swedish nouns are compounded by concatenat-

[1] http://sv.wikipedia.org/

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 99–103,
Gothenburg, Sweden, 26-27 April 2014.

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

Table 1: Compound formation in Swedish; adapted from Stymne and Holmqvist (2008).

ing nouns to each other, creating a single unbroken unit. Compound nouns sometimes come with the interfixes *-s* or *-t*, sometimes without the trailing *-e* or *-a* from the first compounding noun, and sometimes a combination of the two. It should be noted that this is not an exhaustive list of interfixes, there are some other, more specific rules for noun compounding, justified by for example orthographic conventions, not included in Table 1, nor covered by the splitting algorithm. Table 1, adapted from Stymne and Holmqvist (2008), shows the more common modifications and their combinations.

In Koehn and Knight (2003) an algorithm for splitting compound nouns is described. The algorithm works by iterating over potential split points for all tokens of an input corpus. The geometrical mean of the frequencies of the potential constituents are then used to evaluate whether the token split actually is a compound noun or not.

3 Paraphrasing Compound Nouns

To extract candidate compound nouns for paraphrasing, we first tagged the Swedish Europarl corpus and a subset of Swedish Wikipedia using TnT (Brants, 2000) trained on the Stockholm-Umeå Corpus. The resulting corpus was used to compile a frequency dictionary and a tag dictionary, which were given as input to a modified version of the splitting algorithm from Koehn and Knight (2003), producing a list of nouns with possible split points and the constituents and their tags, if any, sorted by descending frequency. The modifications to the splitting algorithm include a lower bound, ignoring all tokens shorter than 6 characters in the corpus. This length restriction is added with the intention of removing noise and lowering running time. Another constraint added is not to consider substrings shorter than 3 characters. The third and last change to the algorithm is the addition of a length similarity bias heuristic to decide between possible split points when there are multiple candidates with a similar result, giving a higher score to a split point that generates substrings which are more similar in length.

Due to the construction of the splitting algorithm, not all split nouns are noun compounds, and without any gold standard to verify against, a set of 200 compound nouns were manually selected by choosing the top 200 valid compounds from the frequency-sorted list. The split compound nouns were then paraphrased by a native speaker of Swedish and validated by two other native speakers of Swedish. The paraphrases were required to be *exhaustive* (not leave out important semantic information), *precise* (not include irrelevant information), and *standardized* (not deviate from other paraphrases in terms of structure).

Nakov and Hearst (2013) have shown that verbal paraphrases are superior to the more sparse prepositional paraphrases, but also that prepositional paraphrases are more efficient for machine translation experiments. However, when examining the compound nouns closely it becomes obvious that the potential paraphrases fall in one of the following four categories. The first category is compound nouns that are easy to paraphrase by a prepositional phrase only, (Examples 1a, 1b), sometimes with several possible prepositions, as in the latter case.

(1) a. psalmförfattare (hymn writer)
författare av psalmer
writer of hymns
b. järnvägsstation (railway station)
station {för, på, längs} järnväg
station {for, on, along} railway

The second category overlaps somewhat with the first category in that the compound nouns could be paraphrased using only a prepositional phrase, but some meaning is undoubtedly lost in doing so. As such, the more suitable paraphrases contain both prepositional and verb phrases (Examples 2a, 2b).

(2) a. barnskådespelare (child actor)
skådespelare som är barn
actor who is child

b. studioalbum (studio album)
album inspelat i en studio
album recorded in a studio

The third and fourth category represent noun compounds that are not necessarily decomposable into their constituents. Noun compounds in the third category can be paraphrased with some difficulty using prepositional phrases, verb phrases as well as deeper knowledge of the semantics and pragmatics of Swedish (Examples 3a, 3b).

(3) a. världskrig (world war)
krig som drabbar hela världen
war that affects whole world
b. längdskidåkning (cross-country skiing)
skidåkning på plan mark
skiing on level ground

Noun compounds in the fourth category are even harder, if not impossible to paraphrase. The meaning of compound nouns that fall into this category cannot be extracted from the constituents, or the meaning has been obscured over time (Examples 4a, 4b). There is no use paraphrasing these compound nouns, and as such they are left out.

(4) a. stadsrättighet (city rights)
b. domkyrka (cathedral)

All compound nouns that are decomposable into their constituents were paraphrased according to the criteria listed above as far as possible.

4 Machine Translation Experiments

To evaluate the effect of compound paraphrasing, a phrase-based statistical machine translation system was trained on a subset of roughly 55,000 sentences from Swedish-English Europarl, with the Swedish compound nouns paraphrased before training. The system was trained using Moses (Koehn et al., 2007) with default settings, using a 5-gram language model created from the English side of the training corpus using SRILM (Stolcke, 2002). A test set was paraphrased in the same way and run through the decoder. We tested two versions of the system, one where all 200 paraphrases were used, and one where only the paraphrases in the first two categories (transparent prepositional and verb phrases) were used. As a baseline, we used a system trained with the same settings on the unmodified training corpus and applied to the unmodified test corpus.

The systems were evaluated in two ways. First, we computed standard BLEU scores. Secondly, the translation of paraphrased compounds was manually evaluated, by the author, in a random sample of 100 sentences containing one or more of the paraphrased compounds. Since the two paraphrase systems used different paraphrase sets, the manual evaluation was performed on two different samples, in both cases comparing to the baseline system. The results are shown in Table 2.

Looking first at the BLEU scores, we see that there is a small drop for both paraphrase systems. This drop in performance is most certainly a side effect of the design of the paraphrasing script. There is a certain crudeness in how inflections are handled resulting in sentences that may be ungrammatical, albeit only slightly. Inflections in the compounding nouns is retained. However, in paraphrases of category 2 and 3, the verbs are always in the present tense, as deriving the tense from the context can be hard to do with enough precision to make it worthwhile. Consequently, the slightly better score for the system that only uses paraphrases of category 1 and 2 is probably just due to the fact that fewer compounds are paraphrased with verbal paraphrases.

Turning to the manual evaluation, we see first of all that the baseline does a decent job translating the compound nouns, with 88/100 correct translations in the first sample and 81/100 in the second sample. Nevertheless, both paraphrase systems achieve slightly higher scores. The system using all paraphrases improves from 88 to 93, and the system that only uses the transparent paraphrases improves from 81 and 90. Neither of these differences is statistically significant, however. McNemar's test (McNemar, 1947) gives a p value of 0.23 for S1 and 0.11 for S2. So, even if it is likely that the paraphrase systems can improve the quality of compound translation, despite a drop in the overall BLEU score, a larger sample would be needed to fully verify this.

5 Discussion

The results from both the automatic and the manual evaluation are inconclusive. On the one hand, overall translation quality, as measured by BLEU, is lowered, if only slightly. On the other, the manual evaluation shows that, for the paraphrased

System	BLEU	Comp	
		S1	S2
Baseline	26.63	88	81
All paraphrases	26.50	93	–
Paraphrases 1–2	26.59	–	90

Table 2: Experimental results. Comp = translation of compounds; S1 = sample 1; S2 = sample 2.

compound nouns, the experimental decoders perform better than the baseline. However, this improvement cannot be established to be statistically significant. This does not necessarily mean that paraphrasing as a general concept is flawed in terms of translation quality, but judging from these preliminary results, further experiments with paraphrasing compound nouns need to address a few issues.

The lack of quality in the paraphrases, probably attributable to how inflections are handled in the paraphrasing scripts, might be the reason why the first experimental system performs worse than the second. This could indicate that there is little to be won in paraphrasing more complex compound nouns. Another possible explanation lies in the corpus. The tone in the Europarl corpus is very formal, and this is not necessarily the case with the more complex paraphrases.

The number of compound nouns actually paraphrased might also attribute to the less than stellar results. If, when training the experimental systems using the paraphrased Swedish corpora, the number of non-paraphrased compound nouns outweigh the number of paraphrased compound nouns the impact of the paraphrases might actually only distort the translation models. This could very well be the problem here, and it is hard from these experiments to judge whether or not the solution is to have more paraphrasing, or none at all.

6 Conclusion

We have reported a pilot study on using paraphrasing of compound nouns to improve the quality of machine translation from Swedish to English, building on previous work by Nakov and Hearst (2013). The experimental results are inconclusive, but there is at least weak evidence that this technique may improve translation quality specifically for compounds, although it may have a negative effect on other aspects of the translation. Further experiments could shed some light on this.

There are a couple of routes that are interesting to follow from here. In Nakov and Hearst (2013), a number of verbal and prepositional paraphrases are gathered through the means of crowd sourcing, and compared to paraphrases gathered from a simple wild card keyword search using a web based search engine. Since the paraphrases in the experiments described in this paper are done by the author and verified by no more than two other native speakers of Swedish, the paraphrases might not be generic enough. By crowd sourcing paraphrase candidates the impact of one individual's personal style and tone can be mitigated.

Another interesting topic for further research is the one of automated compound noun detection. The algorithm used for splitting compound nouns returns a confidence score which is based on the geometrical mean of the frequencies of the constituents together with some heuristics based on things such as relative length of the constituents and whether or not the constituent was found at all in the corpus. This confidence score could potentially be used for ranking not the most frequently occurring compound nouns, but the compounds where the classifier is most confident.

A number of improvements on the applied system can probably lead to a wider coverage. For one, to alter the algorithm so as to allow for recursive splitting would help in detecting and disambiguating compound nouns consisting of three or more constituents. This might be helpful since, as previously mentioned, Swedish is a highly productive language, and it is quite common to see compound nouns consisting of three or more constituents. It should be noted however, that for this to have the desired effect, the paraphrasing would have to be done recursively as well. This could potentially lead to very long sentences generated from very short ones, if the sentence includes a compound consisting of three or more parts.

Some other minor improvements or possible extensions over the current implementation includes taking into account all orthographical irregularities to get a broader coverage, running the algorithm over a more domain specific corpus to get more relevant results, and finally, automating the actual paraphrasing. This last step, however, is of course far from trivial.

References

Thorsten Brants. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 224–231.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the Semantics of Noun Compounds. *Computer Speech & Language*, 19(4):479–496.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics*, pages 187–193.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, and Richard Zens. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157.

Preslav I. Nakov and Marti A. Hearst. 2013. Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases. *ACM Transactions on Speech and Language Processing*, 10(3):1–51.

Jonas Sjöbergh and Viggo Kann. 2004. Finding the correct interpretation of swedish compounds, a statistical approach. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Andreas Stolcke. 2002. SRILM: An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*.

Sara Stymne and Maria Holmqvist. 2008. Processing of swedish compounds for phrase-based statistical machine translation. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, pages 180–189.

Sara Stymne. 2009. *Compound Processing for Phrase-Based Statistical Machine Translation*. Ph.D. thesis, Department of Computer and Information Science, Linköpings Univ.