Paraphrasing of Swedish Compound Nouns

Edvin Ullman
Department of Linguistics and Philology, Uppsala University
edvinu@stp.lingfil.uu.se

Abstract

The goal for this project is to examine and evaluate the effect of paraphrasing noun-noun compounds, with the aim of improving machine translation. The paraphrases will 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 are inconclusive, with no evidence of the use of, or damage of, paraphrasing Swedish compound nouns in relation to machine translation.

1 Credits

This paper was made possible by the grace of my dear fellows at the Institution for Linguistics and Philology, Uppsala University, who, despite my situation outside of academia, have shown nothing but support and patience.

2 Introduction

Swedish, together with many other Germanic languages, is a highly productive language in the sense that 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).

The assumption that paraphrases of noun compounds can help in machine translation is backed 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 verbal paraphrases, relying on native speaker intuition for the quality and correctness of the paraphrases.

2.1 Related Work

Studies in theoretical linguistics 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 holds 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 (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

1http://sv.wikipedia.org/
form a third, two big focal points in previous studies has been detecting compound nouns (Sjöbergh and Kann, 2004) and splitting compound nouns (Stymne, 2008 and 2009).

2.2 Swedish Compound Nouns

Swedish nouns are compounded by concatenating nouns to each other, creating a white space delimited, unbroken unit. Compound nouns sometimes come with the interfixes -s or -t, sometimes the trailing -e or -a from the first compounding noun, and sometimes a combination of the two. There are some other, more specific rules for noun compounding, justified by for example orthographic convention. Table 1 shows the more common modifications and their combinations. These modifications are the ones used for the splitting algorithm. The table is, with the exception of the excluded modifications, borrowed from (2008).

The splitting algorithm is a modification of (Koehn and Knight, 2003), which works by iterating over potential split points for each noun token of at least a certain length in the corpus. This length restriction is a restriction added from the original algorithm with the purpose of removing noise and increasing performance. Another restraint is added to not consider substrings of a length shorter than three. The geometrical mean of the frequencies of the two substrings in a frequency dictionary compiled from a subset of Swedish Wikipedia is used to determine which split point is the more likely. The third and last change to the algorithm is the addition of a length similarity bias heuristic with the purpose of aiding in deciding between possible split points when there are multiple candidates with a similar or near similar result, giving a higher score to a split point that generates substrings which are more similar in length.

2.3 Paraphrasing Compound Nouns

Due to the construction of the algorithm, not all split nouns are noun compounds, and without any golden standard to verify against, a selection of about 200 compounds were considered for paraphrasing. This selection represents the most frequently occurring noun compounds in the Swedish EUROPARL corpus and a subset of Swedish Wikipedia. The split compounds are then paraphrased by a native speaker of Swedish and validated by two other native speakers of Swedish. The following criteria would intuitively constitute a good paraphrase:

- Exhaustive – a paraphrase should explain what relationship holds between its constituting parts, not leaving out important semantic information,
- Precise – a paraphrase should not include information that is not present in a given definition of its compounds,
- Standardized – a paraphrase should not deviate too far from other paraphrases in terms of structure, taking care not to include too specific or localized word forms or tenses

(2013) has 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 within one of four categories.

The first category is compound nouns that are easy to paraphrase by a prepositional phrase only. For some, multiple prepositions are fit to use in the paraphrase.

- psalmförfattare
  författare av psalmer

- järnvägsstation
  station {för, på, längs} järnväg

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 verbal and prepositional phrases.

- barnskådespelare
  skådespelare som är barn

- studioalbum
  album inspelat i en studio

Not all noun compounds are necessarily decomposable into its constituents. These compound
nouns can broadly be divided into two categories. The first category of compound nouns that can be paraphrased with some difficulty using prepositional phrases, verbal phrases as well as deeper knowledge of the semantics and pragmatics of Swedish.

- världskrig
  krig som drabbar hela världen

- längdskidåkning
  skidåkning på plan mark

The second category is 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. There is no use paraphrasing these compound nouns, and as such they are left out.

- stadsrättighet
- domkyrka

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

Evaluation is done by training a decoder in Moses, with the Swedish compound nouns paraphrased before training. This is compared against a baseline decoder, trained on the unmodified parallel corpus. The translations are scored using BLEU scores.

### 3 System Description

For splitting nouns into constituents, the Swedish EUROPARL corpus and the subset of Swedish Wikipedia was tagged using TnT (Brants, 2000). The resulting corpus is used for compiling a frequency dictionary and a tag dictionary. These two files are used with a splitting algorithm which is a modification of (Stymne and Holmqvist, 2008). The resulting file contains a list of nouns with possible split points and the constituents and their tags, if any, sorted in descending frequency.

For evaluating, a 5-gram language model was created from a subset, consisting of roughly 55,000 sentences of the EUROPARL corpus using SRILM (Stolcke, 2002), and trained in the Moses tool-kit (Koehn et al., 2007). This constitutes the baseline decoder against which the results from the experimental decoders will be compared.

A simple script was run to replace instances of the paraphrasable noun compounds with their paraphrases in the Swedish corpus. A language model was then trained with this altered corpus, and an experimental decoder was trained, again using Moses.

### 3.1 Models

The models are all trained on a parallel corpora of roughly 55,000 sentences from the EUROPARL corpus. Training and translating on a larger corpus was desirable, but due to time constraints and build times this smaller subset suffices.

<table>
<thead>
<tr>
<th>System</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish baseline</td>
<td>1072158</td>
</tr>
<tr>
<td>paraphrased</td>
<td>1096279</td>
</tr>
<tr>
<td>English baseline</td>
<td>1185507</td>
</tr>
</tbody>
</table>

Table 2: Corpora size.

As shown in Table 2, the paraphrased Swedish corpus is, with 24,121 more tokens, only about 2
percent larger than its baseline counterpart. This is due to the manual procedure included in the paraphrasing and the limited time to perform it. A way of increasing the impact of paraphrasing would be to extract the n most frequent compound nouns from the training data. This would however result in a loss of coverage, as more general compound nouns extracted from a more generalized corpus most certainly are not frequent enough to be extracted from the training data alone. Another way would be to apply some form of paraphrasing logic to automatically paraphrase identified compound nouns, with the risk of making incorrect splitting and paraphrasing.

4 Results

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish</td>
<td>English</td>
<td>26.63</td>
</tr>
<tr>
<td>English</td>
<td>Swedish</td>
<td>20.84</td>
</tr>
</tbody>
</table>

Table 3: Results from the baseline decoder.

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish</td>
<td>English</td>
<td>25.77</td>
</tr>
<tr>
<td>English</td>
<td>Swedish</td>
<td>19.71</td>
</tr>
</tbody>
</table>

Table 4: Results from the experimental decoder.

When paraphrasing the Swedish corpus, the performance of the decoder drops about 1 point both directions. This lowered performance may be the result of a number of different reasons.

For one, the script used for paraphrasing compound nouns can only handle inflections so well. Some slight distortion of the corpus was unavoidable with the current implementation. This could well be improved upon by either excluding all but the most simple paraphrases. Another approach would be to use a more elaborate script.

If a wide coverage is desirable, and all levels of complexity of compound nouns should be covered, then maybe another approach of obtaining suitable paraphrases should be applied. Crowd sourcing paraphrase candidates is a method that comes to mind for this task.

To assess wherein the problems lie, another paraphrase script was written, roughly half in size of the original and comprised only of simple prepositional paraphrases. This was then used to paraphrase the Swedish training data, and a new model was trained. The resulting BLEU scores can be found in Table 5.

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish</td>
<td>English</td>
<td>26.22</td>
</tr>
<tr>
<td>English</td>
<td>Swedish</td>
<td>20.46</td>
</tr>
</tbody>
</table>

Table 5: Results from the second experimental decoder

Further experiments were conducted with the decoders. The paraphrasing script was used to preprocess the testing data, and then fed to the three decoders. The results can be seen in Table 6.

<table>
<thead>
<tr>
<th>Decoder</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.59</td>
</tr>
<tr>
<td>Experimental 1</td>
<td>26.50</td>
</tr>
<tr>
<td>Experimental 2</td>
<td>26.01</td>
</tr>
</tbody>
</table>

Table 6: Paraphrased Swedish test data to English.

4.1 Discussion

In the summary in Table 7, the scores from all experiments can be viewed. The best obtained BLEU scores are in bold, and as shown, the experimental decoders largely perform worse than the baseline decoder. This does not necessarily mean that paraphrasing as a general concept is flawed in terms of decoding quality to and from Swedish, but judging from these preliminary results, further experiments with paraphrasing compound nouns need to address a few issues.

The experimental decoders perform almost as well as the baseline decoder when translating from their respective Swedish corpus. In other words, the performance of the decoder relies, in part, on whether or not the testing data is paraphrased to the same extent. This might imply that the quality of the paraphrases is lacking, and that the only way to cover for this is to use a decoder trained on the same, poorly paraphrased corpus. The lack of quality in the paraphrases might lie in how inflections are handled in the paraphrasing scripts.

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. Since the paraphrases 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, this can be avoided.

The number of compound nouns actually paraphrased might also contribute to the less than stellar results. If, when training the experimental decoders using the paraphrased Swedish corpora, the number of non paraphrased compound nouns outweigh the number of paraphrased compound nouns the impact might 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.

5 Conclusion

The fact that the experimental decoders perform best with paraphrased testing data is a point of interest. This could, as has been covered in the discussion, mean on of two things. Either it is a sign of poor quality of the paraphrases, and it very well may be, or it is not. Regardless of which, the performance when paraphrasing the test data is all but par with the baseline decoder. If a decoder were to be trained with a Swedish corpus with even more noun compounds paraphrased, the impact of paraphrasing the test data might surpass the overall lowering of performance, resulting in higher performance after all.

From the data obtained it is hard to determine the full use of paraphrasing Swedish compound nouns. Further experiments could shed some light on this.

5.1 Ideas for Further Research

There are a couple of routes that are interesting to follow from here. In (2013), a number of verbal and prepositional paraphrases is 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. The accuracy of the paraphrases would probably be better with this approach.

Another interesting topic for further research is the one of automated compound noun detection. The algorithm used for splitting compound nouns return a certainty 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 certainty score could potentially be used for ranking not the most frequently occurring compound nouns, but the most guaranteed compound nouns.

A number of improvement 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 would be very 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.

Some other small 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, might actually not be considered a small one.

References


Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, USA.


