Paraphrasing compound nouns in Swedish
Project proposal
Edvin Ullman

**Purpose and aims**
The purpose of this research project is to replicate, and adapt for Swedish, previous experiments done on paraphrasing compound nouns, with the help of verbs and prepositions, in order to improve statistical machine translations. This serves not only the purpose of improving translations, but also to explore the semantic relationship between the nouns that make up Swedish compound nouns.

The aim is to establish a database over Swedish compound nouns and their paraphrased counterparts, to enable further investigation and experimentation. The use for this database will be evaluated and hopefully justified by experimentations in statistical machine translation. In previous studies with paraphrased compound nouns in English, this has in fact been shown to improve the results.

**Survey of the field**
Interpretation of multiword expressions, spanning verb-particle constructs, compound nouns, idiomatic expressions, and more still, is a broadly investigated field within language technology. In this project I will be narrowing the scope to only contain Swedish compound nouns.

There are no studies on paraphrasing Swedish compound nouns, but rather of how to correctly identify and split Swedish compound nouns. As Swedish and other germanic languages largely share the way compound nouns are constructed, namely through concatenation of separate nouns, it is reasonable to consider studies done on those languages as well. However, even with this expanded scope, studies on paraphrasing compound nouns seem to be nonexistent.

In Nakov & Hearst (2013), tests done in statistical machine translation with paraphrased compound nouns yielded an increased BLEU score to the equivalent of 33% to 50% compared to doubling the training data, and there’s reason to believe that this result could be replicated for other languages as well.

**Programme description**
The idea is to replicate the method used in Nakov & Hearst (2013), and Nakov (2008) where they used Amazon Mechanical Turk to establish the most probable, or fitting, paraphrases for a number of English noun-noun compounds. To be able to get this project done within the assigned timeframe, the crowd sourcing method I will be using might differ from the previously mentioned, besides, although cheap, Amazon Mechanical Turk comes at a cost.

To test if the acquired paraphrases help with statistical machine translation, the paraphrased compound nouns will be inserted in place of their non-paraphrased counterparts, and a suitable statistical machine translation algorithm will be trained using the sentences generated this way.

The first step is as previously mentioned the gathering of information. This should be a dynamic process, and there is no need to specify a particular end time for this part. After all, establishing an appropriately sized database of paraphrased noun compounds for Swedish is...
in itself a valuable task. That said, enough data needs to be gathered to allow the project to continue.

The second step is to generate paraphrased compound nouns and substitute existing compound nouns with those paraphrases in a suitable corpus. The method for doing this will simply be scripted and automated. As such, the actual work time at this step is hard to estimate at this time. It can however be worked on in parallel with step one and should as such be of no concern.

The third step is maybe the biggest step. Depending on the corpus and the choice of statistical machine translation algorithm, there may be need for implementation at this step. Regardless, however, training and tuning needs to be done, and finally, the testing itself. This step needs the two preceding steps to be done, or in the case of step one, done enough.

The fourth step is to evaluate whether or not paraphrasing compound nouns in Swedish aids in statistical machine translation. This is what it all boils down to. BLEU scores will be computed for the translation for incrementally sized training data, with and without paraphrased compound nouns.

Significance
There exists a need for deeper understanding of the interpretation of multiword expressions in language technology, and compound nouns in Swedish are no exception. As there exists no paraphrased compound noun database, the significance of establishing one is obvious. More so, Swedish, like many germanic languages, has particularly troublesome compound nouns, as compound nouns in Swedish are one blank-separated entities, making them even more cumbersome to handle in most language technology applications.

Equipment
As has been mentioned, an attractive alternative for crowd sourcing the information needed is to use Amazon Mechanical Turk. But in the event that this is not the suitable, a number of alternatives exist.

Google Docs provide a collaborative solution for surveys. Although time consuming to design, the ease of distribution through social media makes for a very attractive alternative.

Designing a survey tool bottom-up, and host it either through the University or privately. There are many advantages of doing this, but for obvious reasons this might be even more time consuming than creating a survey in Google Docs.

All these methods of course suffer from the human factor. There will be bad data, and it's hard to determine beforehand the scope of this.

References


Paraphrasing compound nouns in Swedish
Project proposal
Edvin Ullman
