Extracting Polish translations of German compound nouns from a large bilingual corpus

Marcin Junczys-Dowmunt (Adam Mickiewicz University, Poznań, Poland)

In this talk an attempt will be made to demonstrate how statistical word alignment systems can be used for linguistic knowledge extraction from bilingual parallel corpora.

We focus on statistical methods to identify, align and extract Polish translations of German compound nouns. The collected translation pairs are meant to serve as source and training data for further studies in contrastive linguistics and natural language processing, especially in machine translation.

Consisting of more than 17,000 parallel German-Polish documents extracted from the Official Journal of the European Union dated from May 1st, 2005 to April 30th, 2007, our corpus contains over 19 million tokens for each language. These documents are legal texts, treaties, international agreements, parliamentary questions, etc. and concern for instance agricultural, economical, technical, or political matters. When written in German, they are known for their richness (some may say overuse) of compound nouns. In this corpus we have identified over 80,000 different German compound noun types that constitute approximately half of the vocabulary of the corpus; again nearly half of these compounds are *hapax legomena*. Any natural language processing application for German must be able to deal with this kind of problem, the dimension of which we will illustrate in a brief quantitative analysis of the compounds in the corpus.

The state-of-the-art statistical word alignment system GIZA++ (Och 2000) is used to produce alignments and probability translation tables. Possible translation candidates for the compound nouns we have identified in the German half of the corpus are gathered from the alignment data. The effects of various preprocessing and postprocessing steps performed on the corpus and the received alignment data are discussed. German and Polish, being both highly inflectional languages with significant differences in word order principles, are expected to be hard to align. Therefore we concentrate on methods to increase the alignment precision for noun phrases only. Lemmatization, deletion of function words, splitting of compound words, and other cleaning procedures are performed on the raw corpus data.

We investigate the impact of using linguistically motivated word classes on the alignment process in contrast to automatically obtained word classes (Och 1993) and we evaluate three alignment refinement methods (Och and Ney 2003) on the GIZA++ generated output.

The quality of the obtained alignments is measured against a set of compounds manually aligned with their Polish counterparts. We evaluate alignments of compound tokens within selected sentence pairs instead of compound types since types can have several correct counterparts, which may be hard to capture in a test set. When splitting is used, compounds and their counterparts are aligned to the level of atomic compound segments.

Finding the best translation for compound types among the translations collected and the rejection of incorrect alignments is another problem we plan to address. Our methods include translation probability evaluations based on IBM's Model 4 and checks of grammatical wellformedness and adjacency of the Polish equivalent phrases.

At the end of the talk a short overview of the data gathered is given and we attempt a first comparison to earlier studies (Jeziorski 1982) that were concerned with German compounds and their Polish counterparts, but were not based on large corpora and did not employ automatic methods of data extraction.

References

Jeziorski, Jan. *Substantivische Nominalkomposita des Deutschen und ihre polnischen Entsprechungen.* Wydawnictwo Polskiej Akademii Nauk , Poland 1980.

Och, Franz J. *Maximum-Likelihood-Schätzung von Wortkategorien mit Verfahren der kombinatorischen Optimierung*. Studienarbeit, Universität Erlangen-Nürnberg, Germany 1995.

Och, Franz J. *Giza++: Training of statistical models*. Available at http://www.fjoch.com/GIZA++.html, 2000.

Och, Franz J., Ney Hermann. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, volume 29, number 1, pp. 19-51, March 2003.