Project proposal:

Mapping MWE’s in the Swedish Treebank

1. Purpose and aims

Recent research has pointed out that multiword expressions (MWE) are important to take into account when parsing texts. Many experiments have shown that the accuracy is improving if the parser is being trained with annotated data that have marked up MWE’s. MWE are often being classified as: lexicalized phrases, like fixed expressions that having a fixed word order (e.g. ad hoc), semi-fixed expressions that have fixed word order but can come with some lexical variations (e.g. kick the bucket) and syntactically flexible expressions that have a wide range of syntactic variability (e.g. put up), and the other category is institutionalized phrases that are phrases that are semantically and syntactically compositional, like compounds (e.g. apple juice).

MWE lexicons have been used in various tasks like syntactic parsing and machine translations. There exist monolingual, bilingual and multilingual MWE lexicons, however most of them are specified for specific task or are only present in one particular language. One could also extract MWE’s using corpuses and statistical methods but that requires a widely distributed corpus and corpuses that have the MWE’s marked with a special mark. Many corpuses that exist today have also problem with sparse data to give any adequate results regarding MWE identification and extraction.

There exist some corpuses which have MWE already marked. One is the French Treebank (FTB) [Abeillé et. al. 2003] which is made up of journalistic articles from Le Monde newspaper. One other is the Swedish Treebank (Talbanken05) which consists of manually annotated data based on informative texts from official sources [Einarsson, 1976]. The British National Corpus are annotated with MWE’s but its coverage is far from complete [Shigeto et. al. 2013].

The aim of this project is to build a monolingual MWE lexicon based on extracted data from the Swedish corpus Talbanken05. At present time there exist not any MWE lexicons for the Swedish language.
2. Survey of the field
The latest work regarding building a MWE lexicon is the ongoing project of construction of an English MWE dictionary [Shigeto et. al. 2013]. They notice in their paper that even if there are a variety of WME researches, only a few of them focus on MWE lexicon construction. Earlier work is for example the French adverb dictionary [Laporte et. al. 2008], a Dutch MWE dictionary [Grégoire, 2007], and a Japanese MWE dictionary [Shudo et. al. 2011].

3. Project description
The goal of this project is to construct a Swedish MWE lexicon with an output format in Lexical Markup Framework (LMF) [Francopoulo et. al., 2006]. The choice of the output format is simply to follow a standardized framework for NLP lexicons. It has been used before for similar task of constructing MWE lexicons and also, on the homepage there are samples (mainly from the LIRICS project) of uses for multiword unit patterns together with LMF.

Scientific questions that will be addressed are:

- Are there any inconsistency regarding multiword expressions in the Swedish Treebank?
- Are all the different categories of multiword expressions present or have some category been overlooked?

3.1. Corpus
The corpus that will be used is the Swedish Treebank Talbanken05 as it is the only existing Swedish Treebank with multiword expression annotated in the corpus.

3.2. Process
In the first part of this project will be to identify the different MWE’s present in the corpus and to categorize them. There will surely be some ambivalent data and in these cases one has to make decisions whenever they are to be considered MWE’s or not.

The second part will be developing a tool for transfer the identified MWE’s from the corpus into a lexicon following the Lexical Markup Framework. This will probably take most of the time for the project.

The last part will be examining the extracted data and trying to find the answers on the scientific questions mention above. The first question will probably be answered during the first part of the project as I will already in this stage see if there are some inconsistencies in the data. The second question could be harder to answer as this could mean searching for data that is not available. In this case some external resources might be needed to have some reference material.
4. Significance
As Shudo et. al. (2011) notes in their paper there have been many attempts to extract MWE’s automatically using statistical corpus-based methods. However the effort it is still difficult to adequately extract MWE’s using a statistical approach as corpuses are not widely distributed. If you consider corpuses in different languages with marked MWE’s there are even fewer. There is a paradox when it comes to compile a MWE lexicon as it is impossible to evaluate the automatic extraction by recall rate without a standard reliable MWE lexicon. To use dictionaries used for human readers are mostly not sufficient as they often lack entries that are frequently encountered in common textual materials. All this points to that there is a need for MWE lexicons.

References


