Identification of Multiword BBN Named Entities in Dependency Annotation of Wall Street Journal

Valletta poster proposal, Working Group 4, work in progress

Eduard Bejček, Pavel Straňák
Charles University in Prague, MFF, ÚFAL
{bejcek,stranak}@ufal.mff.cuni.cz

We have already presented (at PARSEME meeting in Athens, 2014) a method for identifying MWEs from lexicon in dependency structures. We made several experiments only for Czech and we used an existing lexicon of multiword lexemes.

In this work, we want to change the experiment in several conspicuous aspects while keeping its fundamentals intact to confirm the broader usage of our approach. We switch to English, use Wall Street Journal (WSJ) as a source text, compile a new "lexicon" of observed English MWEs, and finally, this time we search for multiword named entities instead of multiword lexemes/phrases, since to the best of our knowledge there is no exhaustive annotation of all MWEs in WSJ.

Everything that is needed is already prepared by many previous projects. Penn Treebank (a phrase structure annotation of WSJ, among others)[1] is provided with BBN entities annotation.[2] Penn Treebank is also part of PCEDT project[3] providing deep syntactic dependency annotation (called tectogrammatical) of WSJ. We use all of these to test our method.

The procedure is as follows in four steps:

1. The WSJ is divided into train and test datasets
2. All BBN named entities (NEs) are extracted from the train dataset and the lexicon of these NEs is compiled. Only multiword NEs are taken into account. Each NE entry in the lexicon contains also the information about its dependency structure as it was observed in the tectogrammatical annotation of WSJ.
3. This lexicon is used to identify known NEs in the test dataset of WSJ. The intrinsic property of this method is that it cannot identify new, previously unobserved NE.
4. Results are evaluated. Out-of-vocabulary NEs are necessarily a part of all *false negatives*.[4] More important would be the analysis of other mistakes: why some NEs were not found? Further, we investigate, whether several false *positives* could not be omissions in BBN annotation.

There are two goals of the task:

- to test the method that searches for MWEs using information about tectogrammatical structure and possibly confirm its adequacy for different type of MWEs and different language and
- to check the consistency of BBN entity annotation (and add missing multiword named entities, if this is the case).

As a follow-up, the quality and the properties of the translation of NEs from English to Czech can be tested: PCEDT is a parallel treebank with the manual translation of WSJ into Czech. Therefore all BBN annotations can be projected onto the Czech part. Then it is possible to conduct the same experiment here: compile

[1] http://www.cis.upenn.edu/~treebank/

[2] BBN Pronoun Coreference and Entity Type Corpus, LDC2005T33

[3] http://ufal.mff.cuni.cz/pcedt2.0

[4] Cross-validation will be used to eliminate extreme cases of out-of-vocabulary NEs.

Czech lexicon of multiword NEs from train dataset, identify them in the test dataset of the Czech parallel data and compare the results with the BBN annotations projected into the Czech test dataset. Differences between the same experiments on English and Czech side should be further examined.