Corpus disambiguation – beyond the morphosyntax

Viktor Nagy

Research Institute for Linguistics, Budapest
nagyv@nytud.hu

Keywords: supervised statistical word sense disambiguation, decision lists

The Hungarian National Corpus is a large, morphological annotated corpus. Every word has an annotation which contains its lemma, part of speech and morphosyntactic category, produced by an automatic stochastic tagging procedure. This procedure can not distinguish phenomena having the same morphosyntactic distribution but different meaning because it treats the context as a sequence of morphosyntactic categories abstracted from the particular words. Such an analysis is often not adequate to determine which lexeme the word belongs to, because the (lemma, part of speech) pair can assign more than one lexeme to it if they are homographs with the same part of speech (according to the principles of Hungarian lexicography). While the morphosyntactic tagging can disambiguate the two meanings of word form *tűz*, namely *tűz*[1] verb 'to pin' and *tűz*[2] noun 'fire', it can not disambiguate the meanings of *ül* verb (1 'to sit', 2 'to celebrate') or *barát* noun (1 'friend', 2 'monk'). An additional problem arises when a word form belongs to different lemmas with the same syntactic behaviour. The word form *sejtette* can be the definite past singular 3th person form of *sejt* verb 'to guess' and *sejtet* verb 'to let guess', the *lappal* can be the singular instrumental form of *lap* noun 'sheet' and *lapp* adjective/noun 'Lapp'. In order to overcome the problem caused by the boundaries of morphosyntactic tagging, it is necessary to use a sort of word sense disambiguation.

Our approach uses a supervised learning algorithm based on a manually sense-tagged corpus. This corpus consists of sample concordances of the studied ambigous cases, about 200 occurrences for each case. The following context features was taken into account:

- bag of lemmas in a wider context,
- bag of word forms in a narrower context,
- bag of inflectional categories in a narrower context,
- form of the ambigous word.

We evaluated two learning algorithm on our data: the Naive Bayes method and the decision lists method. The studied examples were chosen to cover broad ranges of lexical ambiguity types. The two methods attained the same performance, they reached a precision of about 83% on the average.