Statistical Named Entity recognition for Hungarian

Richard Farkas[1], György Szarvas[1]

[1] Hungarian Academy of Sciences, University of Szeged;
Research Group on Artificial Intelligence,
6720 Szeged, Aradi vértanuk tere 1., Hungary,
{rfarkas, szarvas}@inf.u-szeged.hu

Abstract: In this paper, we present decision tree based statistical Named Entity recognizer system for Hungarian. The model was trained and tested on a segment of the Szeged Corpus, containing short business news articles, collected from MTI (Hungarian News Agency, www.mti.hu). We applied C4.5 for classificaton, and examined the accuracy of the system using training sets of different sizes. For this task we used only numerically encodable information (we excluded the word form itself), which contained some orthographical rules specific to Hungarian, but we trained for the recognition of foreign language proper nouns appearing frequently in business news as well. During the experiments the best results showed an accuracy of 89.6% F measure.

The feature set we used:

- Part of Speech code (for the particular word and for its +/- 4 words context)
- Case code
- Type of the word's first letter (for the particular word and for its +/- 4 words context)
- Contains digit (inside the word form)
- Contains capital letter (inside the word form)
- Contains hyphen (inside the word form)
- Is the word the beginning of a sentence
- Is the word between quotation marks is the sentence
- Word length
- Is the word Arabic or a Roman number
- The quotient of lower case frequency and "ignore case" frequency from Szószablya [4] term frequency dictionary for Hungarian
- The quotient of mid-sentence upper case frequency and all uppercase frequency from Szószablya
- Does one of our dictionaries contain the word (we used dictionaries containing city names, country names, surnames, company types, geographical name endings, stop words (used frequently in lower case inside NEs)) (for the particular word and for its +/- 4 words context)