Natural language preprocessing on Hungarian language texts

Miháczi András, Németh László, Rácz Miklós

Mihaczi.Andras.Janos@stud.u-szeged.hu
nemethl@gyorsposta.hu
Racz.Miklos@stud.u-szeged.hu

University of Szeged, Department of Informatics

The foundation of natural language processing is plain text. The purpose of preprocessing is to divide the text into digestible units, such as sentences, words, punctuation marks, and other special tokens.

In most cases, the segmentation of sentences is a simple task. The main source of the problem is the interpretation of possible sentence-final punctuation marks, because sometimes it is not the sentence boundary that they represent.

Word boundaries are in most cases unambiguous and easily recognizable since words are separated by whitespaces. However, there are some special cases to be handled, e.g. in the case when the punctuation mark forms an actual part of the word.

The recognition of tokens from open token classes is a closely related problem to the aforementioned ones. These tokens contain special characters (e.g.: comma, dot, hyphen, quotation marks, etc.) or whitespaces.

The recognition of proper names is also a similar problem. Proper name recognition has to be included in the preprocessing phase as well, since it can influence the segmentation of words based on the fact that a proper name may consist of more than one word. The recognition is usually conducted by using dictionaries, and special rules can be defined to build the proper names. (e.g.: person name = first name + last name)

To solve the above described problem, we tried different methods using the Szeged Corpus (an annotated corpus of 1.2 million words developed by the University of Szeged, Department of Informatics and MorphoLogic Ltd). For sentence and word segmentation we have applied decision tree learning algorithms. The aim of learning was to decide whether punctuation marks form part of the words or mark the end of a sentence. The recognition of proper names was aided by large dictionaries containing all proper names from the corpus, and special dictionaries and rules (created and defined by Research Institute for Linguistics at the Hungarian Academy of Sciences) were also used. The results produced by the recognition methods were also tested on the corpus.

In our paper, we present a preprocessing system (forming the initial part of a module chain) developed by ourselves and other existing methods. The preprocessing module recognizes the sentence and word boundaries, special tokens and proper names. The results are passed on to the next module – the morpho-syntactic processing.