More about Words and Parts of Speech (Concerning Natural Language Processing)

Károly Bibok

University of Szeged, Dept. of Russian Philology
Egyetem u. 2. 6722 Szeged, Hungary
kbibok@lit.u-szeged.hu

In 2000–2002 a consortium of the University of Szeged and MorphoLogic Ltd. (Budapest) developed a morpho-syntactically parsed and disambiguated corpus for Hungarian up to one million text words (tokens) without punctuation characters (for Version 1.0 see http://inf.u-szeged.hu/lll). During this project some problems of the natural language processing (NLP) were realized, the complete solution of which could not been attempted. The following can be mentioned: eliminating mistakes made by text segmentation program, treating text words as morpho-syntactic units, and creating word classes for tokens which do not fit into the traditional parts of speech. The current paper deals with these three problematic cases which more attention should be paid to in the future.

For the written form of a language, a text word is a minimal fragment of a text occurring between spaces, not including punctuation marks (Papp 1968: 190). A program that carries out text segmentation task can be based on this classical definition of the text word if it is made more precise with respect to: 1. tokens containing not detachable punctuation marks on one or another side (e.g. abbreviations ending in points) and 2. tokens having punctuation marks in their internal parts (e.g. e-mail addresses).

Some sequences of text words, however, have to be considered grammatical forms of units, lexical or generated by rules of word-formation and coordination. Therefore, the components, i.e. text words, of such units (e.g. of a complex proper noun) should be drawn together before the morpho-syntactic parsing starts.

Furthermore, in case of a very large and stylistically heterogeneous corpus like ours, the traditional set of parts of speech should be extended in order to classify tokens containing special (punctuation) marks ("://", "\", "@" etc.) in the sub-language of computing or tokens consisting of combinations of numbers and punctuation marks in sports news.