A Proposed New Tool Chain for Corpus Statistics and Collocation Search

Balázs Kis, Gábor Ugray

MorphoLogic
{kis,ugray}@morphologic.hu

The preparation of linguistic resources, namely, corpora and lexicons, is the most work-intensive activity in computational linguistics. Still, only a few papers are published on this topic at general conferences – perhaps because in this field it is the most difficult to achieve new scientific results. This paper emphasizes how existing linguistic tools can be used to more efficiently prepare and process corpora.

The paper presents a new tool chain for corpus preparation and statistics that is generally suitable for collocation searches in monolingual corpora, and is able to produce valuable results even from unannotated corpora.

This tool chain covers the entire process from corpus preparation to the evaluation of statistic results.

The most important step in corpus preparation is providing a consistent format in order to make the text easy to process. To this end, the tool chain applies an XML format that facilitates the representation of texts either with or without linguistic annotation. The tool chain is capable of derive this format in a robust fashion, even directly from common file formats such as MS Word documents or RTF files.

Considering collocation search, the second important step is the extraction of collocate candidates from the corpus; later on, these candidates must be evaluated – by means of statistics, for example – to see if they form a real collocation.

Within the proposed tool chain, collocation extraction utilizes language technology to the greatest possible extent. During the extraction process, either bigrams or trigrams can be selected from the text; their components can be identified, and the collocations can be filtered based on their morpho-syntactic or syntactic properties. To achieve this, the tool chain applies one of MorphoLogic's two parser systems, either HumorESK or Moose (the former is preferred for Hungarian, the latter for English). Parsing results – roots of subtrees in parse forests – can form components of bigrams or trigrams.

The extraction process produces greater noise if performed on unannotated corpora. Precision can be significantly improved by first lemmatizing and POS-tagging the text. A POS-tagger module is being integrated with the tool chain at the time of writing the paper.

Noise among the candidates can be significantly reduced by applying a post-processor filtering program, also part of the tool chain. If a part of the noise can be described using simple rules, the misclassified candidates can easily be removed from the data set before counting statistics.