Statistical Extension to the Poliqarp Query Language

Aleksander Buczyński[1]

Abstract

The aim of this paper is to present the statistical extension to the corpus search engine Poliqarp developed at the Institute of Computer Science PAS and currently employed in Polish and Portuguese corpora projects. Poliqarp is a utility for searching large tagged corpora, which features an expressive query language, with support for ambiguous morphosyntactic interpretations and distinction between certain and uncertain information.

So far, Poliqarp has worked as a concordancer, returning a list of matches with contexts of selected width as a response to every query. This provides the user with examples of usage of specific constructions, but one can imagine a number of corpora problems, where browsing through hundreds of occurrences is not so convenient.

The statistical extension to the query language allow the user to ask for the frequency distributions of specified attributes' values – for example, a given word's forms or cases used by a given author – rather than contexts of the occurrences of each match. The extension also provides several statistical measures for collocation detection and investigating correlations between different attributes.

For example, to find different verbs occurring after the word korpus, with an optional adverb between korpus and the verb, one could write:

```
[base=korpus][pos=adv]?[pos=fin] group by -1.base
```

To find bigrams with the highest Symmetric Conditional Probability:

```
[pos!=interp]{2} group by base; 2.base sort by scp
```

'-1.base', 'base' and '2.base' refer to the base forms of the last, first and second segment of the match respectively. Grouping can also be done by orthographic forms, grammatical classes and categories.

The extension has already been implemented and is currently in beta testing. Poliqarp is freely available under the GNU GPL.

[1] Institute of Computer Science, Polish Academy of Sciences
e-mail: olekb@ipipan.waw.pl)