Demo of Langusta – the graph database environment for parsing of Polish

Jan Posiadała, Hubert Czaja, Eliza Szczechla, Paweł Susicki

Scott Tiger S.A.
15 Kolektorska Street, Warsaw, Poland
{janek,czajah,eliza,trzeci}@tiger.com.pl

Abstract

This document briefly describes the basic functionality of the system Langusta, with particular emphasis on the specific characteristics of the presented solution, i.e.. the use of graph database to parse the Polish language with a very strong use of a declarative graph model query language for a graph database (Cypher). In addition, system developers' motivations and inspirations are briefly pointed out as well as its other functionalities different from Polish language parsing.

Keywords: Langusta demo, NLP, graph databases, Cypher, syntax parsing, corpus analysis, written corpora

1. Introduction

The inspiration to undertake the work presented below was a need for a use of working tool to parse sentences in Polish language. Corpora, which would be processed by the desired tool is limited to the class of *edited* texts, which is defined by the following conditions:

- the authors have an internal need for accuracy, clarity and precision,
- the authors write their texts to fulfil their professional duties,
- the authors were educated in writing skills.

Faced with such strong restrictions on processed corpora tends towards the use of purely rule-based parser with hand written evidence rules, which resulted in an attempt to use rule-based parser SPEJD (Buczyński and Przepiórkowski, 2008). Considering all the advantages and disadvantages of using SPEJD as well as (or even primarily) the extension of requirements for projected application resulted in a decision to build the environment for processing on the basis of graph database supporting declarative query language execution. The main element of this work consisted on reimplementation of database engine query language that supports Cypher (Robinson et al., 2014).

2. Langusta

Langusta is consolidated paradigmatically homogeneous environment for processing and analysis corpora of the Polish language, with particular emphasis on syntactic parsing. Langusta stores data in the graph model, and the main way to access to its data is a declarative graph query language Cypher. In addition, parsing rules are expressed in terms of regular Cypher queries. More broadly characteristic of the presented solution, resulting directly from the application of database without scheme is the ease of analysis and cross-check of linguistic hypotheses - in particular the need to create rules for parsing.

2.1 Functionality

Initially database contains the following language resources, represented in the graph model:

- logically compressed Polish morphosyntactic dictionary PoliMorf (Woliński et al., 2012),
- lexical database for the Polish language, plWordNet (Maziarz et al., 2012).

After entering the text from corpus, tokenization takes place and produces a sequence of tokens, which are the basis for the morphosyntactic annotation.

As a result of morphosyntactic annotation, a text structure is formed in the graph database. In this structure for each token there is a corresponding set of nodes representing a collection of interpretations from the morphosyntactic dictionary.

Created structure is ready to start algorithm which applies user-defined set of rules. As a result of algorithm operation, the syntactic trees of sentences from text in parsing are created.
After the end of parsing, the user can - of course by using query language Cypher - precisely analyse the result of parsing, with help of the presentation:

- syntactic tree (forest) for a given sentence,
- the results of parsing for a given token and its surroundings in a sentence,
- the application of the rule in the process of parsing.

References

Buczyński, A. and Przepiórkowski, A. (2008). *Demo: An Open Source Tool for Partial Parsing and Morphosyntactic Disambiguation*. In: *Proceedings of LREC 2008*.

Przepiórkowski, A., Bańko, M., Górski, R.L. and Lewandowska-Tomaszczyk, B., editors. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski A. and Szałkiewicz Ł. (2012) *PoliMorf: a (not so) new open morphological dictionary for Polish*. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 860–864, Istanbul: Turkey: ELRA

Robinson, I., Webber, J. and Eifrem, E. (2013). *Graph Databases*. O'Reilly Media

Maziarz, M., Piasecki, M. and Szpakowicz, S. (2012). *Approaching plWordNet 2.0. Proceedings of the 6th Global Wordnet Conference.* Matsue: Japan.