Slovene NLTK Tagger

Niko Colnerič
Faculty of Computer and
Information Science
University of Ljubljana

Nejc Banič
Faculty of Computer and
Information Science
University of Ljubljana

December 29, 2011

Abstract - NIKICC

This paper describes the process of building Slovene tagger for use in NLTK.

1 Introduction - NIKICC

Tagger processes a sequence of words and attaches a part of speech tag to each word. We implemented tagger for Slovene language in NLTK. NLTK stands for Natural Language Toolkit and is a library for use in Python (currently for version below 3, we used 2.7). NLTK already has some taggers for English and some other languages. Our goal was to implement NLTK tagger for Slovene language. The resault of this project depends on two other larger projects. First is *JOS*[1](in Slovene it stands for: "Jezikoslovno Označevanje Slovenskega jezika"), from which we used the corpuses (large and structured set of texts). Second, *nltk-trainer*[2], is an open-source project hosting on Github, which was used for actuall training of the tagger on Slovene sentences. Our work based mainly on putting all this pieces together into a working Slovene NLTK tagger.

2 JOS Corpus

TODO - BANIČ

3 Nltk-trainer - NIKICC

Nltk-trainer[2] is project, whose author is Jacob Perkins, from which we took the code for building the tagger. We almost exclusively used the script *train_tagger.py*, which has the ability to generate NLTK taggers. Its mandatory argument is corpus in .pos format. Other optional arguments can choose wheather to generate a Sequential Tagger, a Brill Tagger or a Classifier Based Tagger. It also has the possibility to set default tag (the tag for words, for which the tagger doesn't know how to tag) and wheather to evaluate the tagger or not. For evaluation of the tagger (section 6) this argument was used.

4 Taggers description

4.1 Trigram tagger

TODO - NIKICC

4.2 Brill tagger

TODO - BANIČ

4.3 Naive Bayes tagger

TODO - BANIČ

5 Implementation

5.1 Corpus transformation

TODO - BANIČ

5.2 Training the tagger

TODO - NIKICC

5.3 Usage with NLTK

TODO - BANIČ

5.4 Encoding problems

TODO - BANIČ

6 Evaluation

TODO - NIKICC

7 Conclusion

TODO - BANIČ

References

[1] `http://nl.ijs.si/jos/`

[2] `https://github.com/japerk/nltk-trainer`