Slovene NLTK Tagger

Niko Colnerič
Faculty of Computer and
Information Science
University of Ljubljana
niko.colneric@gmail.com

Nejc Banič
Faculty of Computer and
Information Science
University of Ljubljana
MAIL@

ABSTRACT - NIKICC

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

KEYWORDS

I. 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 result 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 actual training of the tagger on Slovene sentences. Our work based mainly on putting all this pieces together into a working Slovene NLTK tagger.

A. Trigram tagger

TODO - NIKICC

B. Brill tagger

TODO - BANIČ

C. Naive Bayes tagger

TODO - BANIČ

II. RELATED WORK

A. 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 **??**) this argument was used.

B. JOS corpora

The basic component for building Slovene tagger is **JOS corpora**. It contains collections of various text in Slovenian language with annotation. For our project we used jos1M corpus that contain 1 million words with it's appropriate lemmas and morphosyntactic descriptions. The JOS is compatible with Slovene *MULTEXT-East morphosyntactic specifications Version 4*[3]. It basic purpose is to define word classes. For each class there are various attributes and their appropriate values, which they can be mapped into morphosyntactic descriptions (i.e. MSD). The structure of JOS corpora is in **Extensible Markup Language (XML)**. We can easily manipulate XML files in various programming languages (e.g. Python with xml.dom.minidom). This is mandatory because *nltk-trainer*[2] expects special form of input file. For further information see III-A.

III. IMPLEMENTATION

A. Corpus transformation

TODO - BANIČ

B. Training the tagger

TODO - NIKICC

C. Usage with NLTK

TODO - BANIČ

D. Encoding problems

TODO - BANIČ

IV. Results

TODO - NIKICC

Average accuracy BrillTagger . Average accuracy NaiveBayes . Average accuracy TrigramTagger .

fraction	*Trigram*	*NaiveBayes*	*Brill*
0.75	0.828648	0.845122	0.849031
0.80	0.830353	0.846063	0.851387
0.82	0.830551	0.846490	0.851527
0.84	0.829855	0.846221	0.850839
0.86	0.830609	0.847159	0.851240
0.88	0.829901	0.847387	0.851680
0.90	0.828807	0.846713	0.850737
0.91	0.827759	0.846211	0.848899
0.92	0.827613	0.846384	0.848964
0.93	0.828121	0.846358	0.849090
0.94	0.827826	0.846064	0.849636
0.95	0.828913	0.847508	0.851041
0.96	0.838798	0.853809	0.860899
0.97	0.840060	0.856799	0.861213
0.98	0.841683	0.858779	0.862869
0.99	0.837875	0.856699	0.857844
average	**0.831711**	**0.848985**	**0.852931**

TABLE I
ACCURACY OF THE TAGGERS FOR DIFFERENT FRACTIONS. FRACTION TELLS PERCENTAGE OF CORPUS USED FOR TRAINING THE TAGGER, THE REST WAS USED FOR EVALUATION.

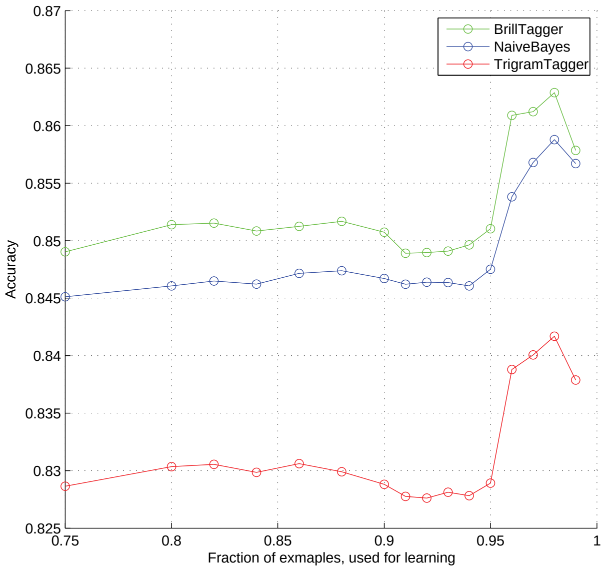

Fig. 1. Evaluation results, with different fractions

V. Conclusion

TODO - BANIČ

Acknowledgements

References

[1] http://nl.ijs.si/jos/
[2] https://github.com/japerk/nltk-trainer
[3] http://nl.ijs.si/ME/V4/msd/html/msd-sl.html