NLP Project: NLTK parser

Juho Kallio

18. joulukuuta 2014

1 Project topic

As a mini-project for the Natural Language Processing course I created a parser with Python and NLTK library for air travel related domain. The program takes as an input the sentence that is tried to be parsed and either returns the parsed structure of it and validates its correctness or tells that the sentence is unparseable. If the sentence is not given as the argument, the program prints out the grammar and the usage instructions. The parser uses a *feature based context free grammar*, shortly FCFG, with 300 production rules.

2 Implemention

I choose to use FCFG because it appeared as the more advanced way to create a context free grammar in the NLTK tutorials. The benefit are the features on the rules that give more expressive power to every production. Apparently, the downsides are limiting ourselves to bottom-up parsing and losing the out of the box sentence generation method that comes with NLTK and standard CFGs. I circumvented this in testing by loading the FCFG as a standard CFG and using that to generate sentences. These sentences were not using the feature side of the rules, so most of them were incorrect. This was, however, enough since now I was able to use the FCFG to parse the sentences and only printed the correct ones.

Test driven development turned out to be a very natural and nice way to create the grammar. I added a new sentence to either to the array of correct or incorrect ones, ran my tests that validated these and when the parser couldn't handle everything correctly, I improved the parser. One limitation of the implementation of the parser is the need to add repetitively all the forms of the words. This could be maybe improved by having a separate morphological unit for generating the right forms of the words, e.g. the right plurals for nouns.

3 Results

The parser focuses on a small domain. Even then, currently the lexicon is primitive. The grammar covers quite many cases, but not nearly everything. The

parse is at its current form mostly a proof of concept and a base that allows quick development of a more advanced FCFG.

4 Instructions to run the program

The program can be downloaded from `https://github.com/juhokallio/airportParser`. It can be run with Python 3 and is tested with the Python version 3.4.0. Running the application requires NLTK (Natural Language Toolkit), which can be downloaded from `http://www.nltk.org/`. The command `python parsing.py` prints out the FCFG and user instructions, the actual parsing happens by running the command `python parsing.py "put a sentence here"`.