Universität des Saarlandes, Fachrichtung 4.7 Allgemeine Linguistik

UNIVERSITÄT DES SAARLANDES

Language Processing for Different Domains and Genres (WS 2009/10)

Daniel C. Müller

Berkley Parser

for detailed information go to http://nlp.cs.berkeley.edu/Main.html#parsing

Grammar

Currently there are the following grammars available:

eng_sm6.gr	English Grammar [for java 1.5]
eng_sm5.gr	English Grammar [for java 1.6]
bul_sm5.gr	Bulgarian Grammar [for java 1.6]
arb_sm5.gr	Arabic Grammar [for java 1.6]
chn_sm5.gr	Chinese Grammar [for java 1.6]
fra_sm5.gr	French Grammar [for java 1.6]
ger_sm5.gr	German Grammar [for java 1.6]

Additionally the parser can learn your own grammar.
You will need a treebank in order to learn new grammars. The package contains code for reading in some of the standard treebanks. To learn a grammar from the Wall Street Journal section of the Penn Treebank, you can execute

```
java -cp berkeleyParser.jar edu.berkeley.nlp.PCFGLA.GrammarTrainer -path <WSJ location> -out <grammar-file>
```

To learn a grammar from trees that are contained in a single file use the -treebank option, e.g.:

```
java -cp berkeleyParser.jar edu.berkeley.nlp.PCFGLA.GrammarTrainer -path <WSJ location> -out <grammar-file> -treebank SINGLEFILE
```

This will read in the WSJ training set and do 6 iterations of split, merge, smooth. An intermediate grammar file will be written to disk once in a while and you can expect the final grammar to be written to <grammar-file> after 15-20 hours. The GrammarTrainer accepts a variety of options which have been set to reasonable default values. Most of the options should be self-explaining and you are encouraged to experiment with them. Note that since EM is a local method each run will produce slightly different results. Furthermore, the default settings prune away rules with probability below a certain threshold, which greatly speeds up the training, but increases the variance. To train grammars on other training sets (e.g. for other languages), consult edu.berkeley.nlp.PCFGLA.Corpus.java and supply the correct language option to the trainer.
To the test the performance of a grammar you can use

```
java -cp berkeleyParser.jar edu.berkeley.nlp.PCFGLA.GrammarTester -path <WSJ location> -in <grammar-file>
```

Input

If you use input files, please use one line for each sentence.

Start

To start the parser you have to use the command:

```
java -jar berkeleyParser.jar -gr <grammar>
```

followed by the input.