NTMI - Project Exercises - Part B

Alex Khawalid (10634207)
Wessel Klijnsma (10172432)
Winand Renkema (10643478)

March 8, 2016

Step 1: Binarization

1 Introduction

The goal of step 1 of part B is to binarize trees modifying the labels in such a way the the proces can be reversed. This is necessary for CKY algorithm which is used in PCFG based parsing.

2 Method

The binarization program consists of 3 steps. First, a tree is convertedto a list containing lists with (tag, word) tuples for each level in the tree. Then this list is binarized and intermediate nodes, with tags that allow for the process to be reversed, are added. Lastly, the binarized tree is converted to a string representation and written to a file.

3 Evaluation

The evaluation results of the PCFG parser with `validate20.txt` and `test20.txt` as input.
`validate20.txt`

```
[Average (up to 725)]  P: 80.17 R: 74.51 F1: 77.24 EX: 23.17
[Average]  P: 80.17 R: 74.51 F1: 77.24 EX: 23.17
```

`test20.txt`:

```
[Average (up to 1034)]  P: 79.61 R: 73.88 F1: 76.63 EX: 20.69
[Average]  P: 79.61 R: 73.88 F1: 76.63 EX: 20.69
```

4 Running the program

The binarization and PCFG parser can be run by invoking the script `test.bash`. Alternatively, the binarization program can be run on its own by using a command of the following format:

```
python b-step1.py -input INPUT_FILE -output OUTPUT_FILE
```