Natuurlijke Taalmodellen & Interfaces
Tejaswini Deoskar
Part B (Step 1)

Lucas van Berkel, 10747958
Ruben Gerritse, 10760326
Gideon Mooijen, 10686290

Report

To start off with the second part of this course, the assignment is to perform a transformation binarization.

Introduction

The goal of *'Binarization'* is to rephrase fragments in the treebank to make them suitable for the next step: *'Markovization'*. Input of the type:

(NP (NNP Rolls-Royce) (NNP Motor) (NNPS Cars) (NNP Inc.))

should be rewritten in the form:

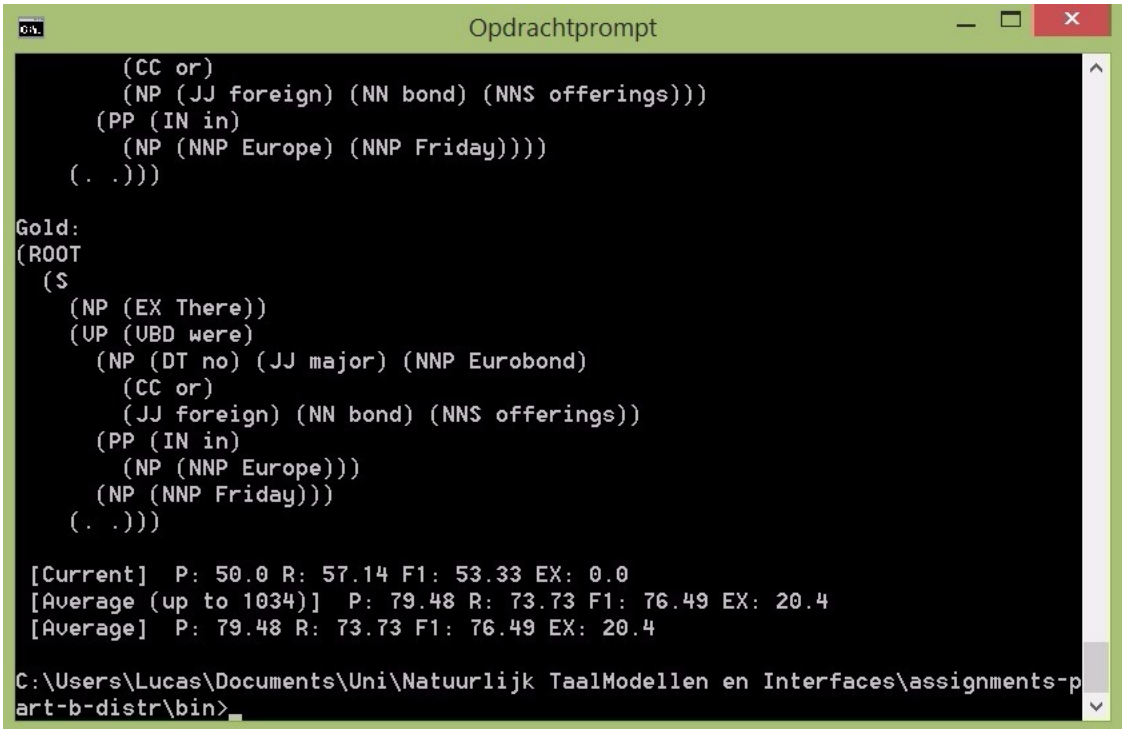
(NP (NNP Rolls-Royce) (@NP→_NNP (NNP Motor)
(@NP→_NNP_NNP (NNPS Cars) (@NP→_NNP_NNP_NNPS (NNP Inc.)))))

Methods

Our code consists of several steps.

1. Convert the textfile to a list
2. Takes all unbinarized sentences and perform step 3 on all of them
3. Binarize one sentence, based on recursion
4. Check: if output of step 2 contains errors
5. Convert binarized sentences to strings, writes them to a text file

Results



```
        (CC or)
        (NP (JJ foreign) (NN bond) (NNS offerings)))
      (PP (IN in)
        (NP (NNP Europe) (NNP Friday))))
    (. .)))

Gold:
(ROOT
  (S
    (NP (EX There))
    (VP (VBD were)
      (NP (DT no) (JJ major) (NNP Eurobond)
        (CC or)
        (JJ foreign) (NN bond) (NNS offerings))
      (PP (IN in)
        (NP (NNP Europe)))
      (NP (NNP Friday)))
    (. .)))

 [Current]  P: 50.0 R: 57.14 F1: 53.33 EX: 0.0
 [Average (up to 1034)]  P: 79.48 R: 73.73 F1: 76.49 EX: 20.4
 [Average]  P: 79.48 R: 73.73 F1: 76.49 EX: 20.4

C:\Users\Lucas\Documents\Uni\Natuurlijk TaalModellen en Interfaces\assignments-p
art-b-distr\bin>_
```

Figure 1: Results of our code

As can be seen in the figure above, the F1-score of our binarization-code is 76.49 %. This is in line with the expectations.

How to run

```
BerkelGerritseMooijen_5.py -input [non-binarized] -output [binarized]
```