ANLP 662 – Homework 4

Ekraam Sabir
USC-ID: 1428-2892-90
esabir@usc.edu

Learning a Grammar

Q1. There are 752 rules in the grammar. Most frequent rule is “PUNC -> .” with a frequency of 346.

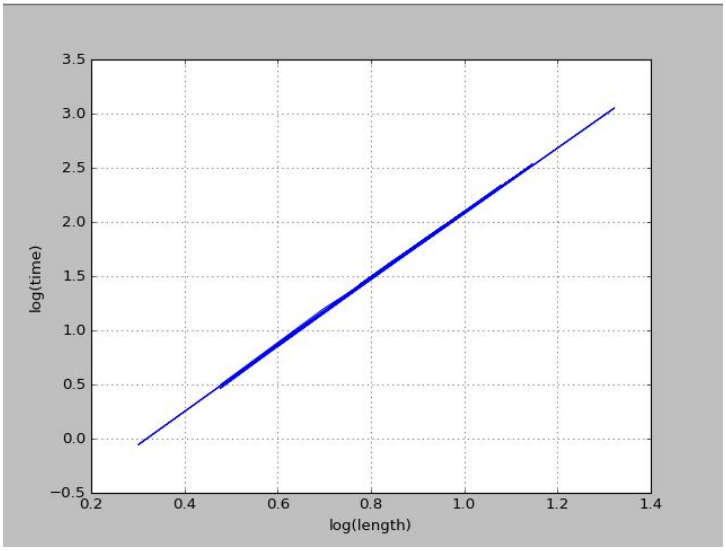
Parsing Sentences

Q2. The output on the first line of dev.strings with log(probability) -18.9464754637 is:

dev.parses:
(TOP (S (NP (DT The) (NN flight)) (VP (MD should) (VP (VB be) (NP (NP* (CD eleven) (RB a.m)) (NP_NN tomorrow))))) (PUNC .))

dev.parses.post:
(TOP (S (NP (DT The) (NN flight)) (VP (MD should) (VP (VB be) (NP (CD eleven) (RB a.m) (NP (NN tomorrow)))))) (PUNC .))

Q3. The value of k is ~3.125 (the slope of the line on a log-log scale). The complexity is supposed to be cubic, which is why it is close to 3. The extra 0.125 is due to the additional constant iterations through the labels for finding X, Y and Z (where X -> Y Z is a rule).



Q4. The output after evaluation from evalb.py is:

dev.parses.post 490 brackets
dev.trees 474 brackets
matching 385 brackets
precision 0.785714285714
recall 0.81223628692
F1 0.798755186722

Improving your Parser

Q5. Modification 1:

I tried extending of smoothing to rules. With the limited number of rules from a small dataset, it is possible that a sentence may never end up with a valid parse. To allow for a complete parse, I added all rules of the form X -> Y Z where (Y and Z are non-terminal labels) with ƛ smoothing (where ƛ=1e-10) and formula (C(X->Y Z) + ƛ)/(N + Bƛ).

dev.parses_smooth.post 496 brackets
dev.trees 474 brackets
matching 408 brackets
precision 0.822580645161
recall 0.860759493671
F1 0.841237113402

Modification 2:

Extended the above smoothing to terminal rules/words with ƛ (ƛ=1e-10). This did not give any improvements. This did not help possibly because the <unk> label already takes care of the distribution of unseen words.

dev.parses_smooth.post 496 brackets
dev.trees 474 brackets
matching 408 brackets
precision 0.822580645161
recall 0.860759493671
F1 0.84123711340

Q6. The output from best parser is:

test.parses_smooth.post 481 brackets
test.trees 471 brackets
matching 404 brackets
precision 0.839916839917
recall 0.857749469214
F1 0.848739495798