Assignment–4

Course Name: Natural Language Processing (COMP 8780)

Course Instructor: Professor Vasile Rus

Submitted By

Student Name: Anjana Tiha
UID: U00619942
Date: 03/15/2018

Problem 1[20 points]

Extract from the BROWN file all grammar rules embedded in parse trees. Do not consider punctuation as a nonterminal. Eliminate numbers attached to non-terminals such as '-1', '-2', etc. Report how many distinct rules you found, what are the 10 most frequent rules regardless of the non-terminal on the left-hand side, and what is the non-terminal with the most alternate rules (i.e. the non-terminal that can have most diverse structures).

Problem 2 [20 points]

Try to estimate how large the above grammar would be if you were to lexicalize it, i.e. to add head words to some of the rules. Work with your own assumptions. The important part for this problem is your general reasoning and not the details. [20 points]

Answer 1 & 2:

Functionality :

Extracted from all the grammar rules embedded in parse trees from the BROWN file. Did not consider punctuation as a non-terminal and Eliminated numbers attached to non-terminals such as '-1', '-2', etc. Reported number of distinct rules, the 10 most frequent rules(regardless of the non-terminal on the left-hand side), and the non-terminal with the most alternate rules.Tried to estimate size of grammar after lexicalizing (adding head words to some of the rules.)

Method :

1. Read BROWN-clean.pos.txt file line by line.
2. Parsed each tree using Top-Down parsing and store the parse tree using custom tree class.
3. Kept list of all the root node for all the sentences.
4. Break down the tree to bring out production/grammar rules for each sentence tree.
5. Stored the unique rules in a python dictionary.
6. Stored right hand side rules in a separate dictionary.
7. Showed the total and distinct grammar size for non lexicalized and lexicalized trees.
8. Showed non terminal with the most alternative rule.
9. Showed 10 most common right side rule.

10. For lexicalization used focused on NP and VP specifically. Used Tree structure to build lexical tree.

11. Grammar rule statistics is reported for complete BROWN corpus file - "BROWN.pos.all".

Performance Report :

Number of distinct rules, the 10 most frequent rules(regardless of the non-terminal on the left-hand side), and the non-terminal with the most alternate rules.

Table 1: 10 most frequent rules(regardless of the non-terminal on the left-hand side) in BROWN.pos.all:

Non-Lexicalized	
Rule	Rule Frequency
IN NP	7016
NP VP	3880
DT NN	3267
NP AUX VP	3047
VBD	1297
IN S	1250
TO	1250
PRP	1232
RB	1224
NNP	1205

Lexicalized	
Rule	Rule Frequency
IN NP	20040
NP VP	10703
NP AUX VP	9411
DT NN	5624
TO	4210
RB	3781
IN S	3692
MD	3641
NNP	3108
VBD	2669

Table 2: Summary for file - "BROWN.pos.all"

		Non Lexicalized	Lexicalized
Total Size of Grammar(Incl Terminal & Non-Terminal)	:	15879	54001
Number of Distinct Grammar rules(Incl Terminal & Non-Terminal)	:	344	1080
Non-terminal(Most Distinct Alternate Rules	:	NP(171)	NP(171)
Number of Distinct Grammar Rules(Regardless of Left)	:	323	1056

Running Instruction :

Script Name: "main.py"
Input : In command line please type: python *main.py*
Data : Extracted "BROWN.pos.all" file.
Output: Output is printed in terminal.