PCFG and Consitituency Parsing

Rajat Nagpal
M.Tech(SE)
rajatnagpal@iisc.ac.in

Github Link to this Report

Abstract

Statistical parsing uses a probabilistic model of syntax in order to assign probabilities to each parse tree. It also provides principled approach to resolve ambiguity in semantics. PCFG, which is the probabilistic version of CFG, can be used to estimate probabilities. CKY parser can be modified for CFG parsing by including in each cell a probability for each non-terminal.

1 Implementation

In this experiment, I have used PCFG method to get the probabilities of the grammar rules. I have used **nltk Penn Treebank** to get the sentences and I have used Induce PCFG library function to get the probabilities after adding **UNK** tokens corresponding to every grammar rule. CKY parser is used after that to get the Parsed Tree. I have calculated **Labelled Precision, Labelled Recall, F1 Score** on the test set.

I have compared my Parser with the following Parser:

http://tomato.banatao.berkeley.edu:8080/parser/parser.html

2 Natural Language Parsing

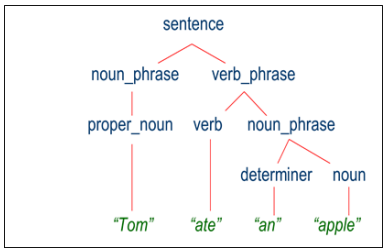
Th analysis of strings of symbols in Natural Language, conforming to the rules of Formal Grammar is called Parsing. It is also referred as Syntax or Semantic analysis of Grammar.

A string of words are formally divided into constituents, then into a Parse Tree, which shows a syntactic relationship among each other and related semantics.

Consider the example *"Tom ate an Apple"*.

The Parse tree will parse the above grammar structure where *sentence* is the root, *verb phase* and *noun phrase* are non terminals and *"Tom"*, *"ate"*, *"an"*, *"apple"* are terminals.

```
sentence -> noun_phrase, verb_phrase
noun_phrase -> proper_noun
noun_phrase -> determiner, noun
verb_phrase -> verb, noun_phrase
proper_noun -> [Tom]
noun -> [apple]
verb -> [ate]
determiner -> [an]
```



3 Probabilistic Context Free Grammar

In formal language theory, a context-free grammar (CFG) is a certain type of formal grammar: a set of production rules that describe all possible strings in a given formal language. Production rules are simple replacements.

PCFG is defined as:

G = (M, T, R, S, P) where

M is the set of non-terminal symbols

T is the set of terminal symbols

R is the set of production rules

S is the start symbol

P is the set of probabilities on production rules

PCFGs extend context-free grammars where each production is assigned a probability. The probability of a derivation (parse) is the product of the probabilities of the productions used in that derivation. These probabilities can be viewed as parameters of the model, and for large problems it is convenient to learn these parameters via machine learning. A probabilistic grammar's validity is constrained by context of its training dataset.

Metrics	Scores
Labelled Precision	63.4
Labelled Recall	63.7
F1 Score	63.56

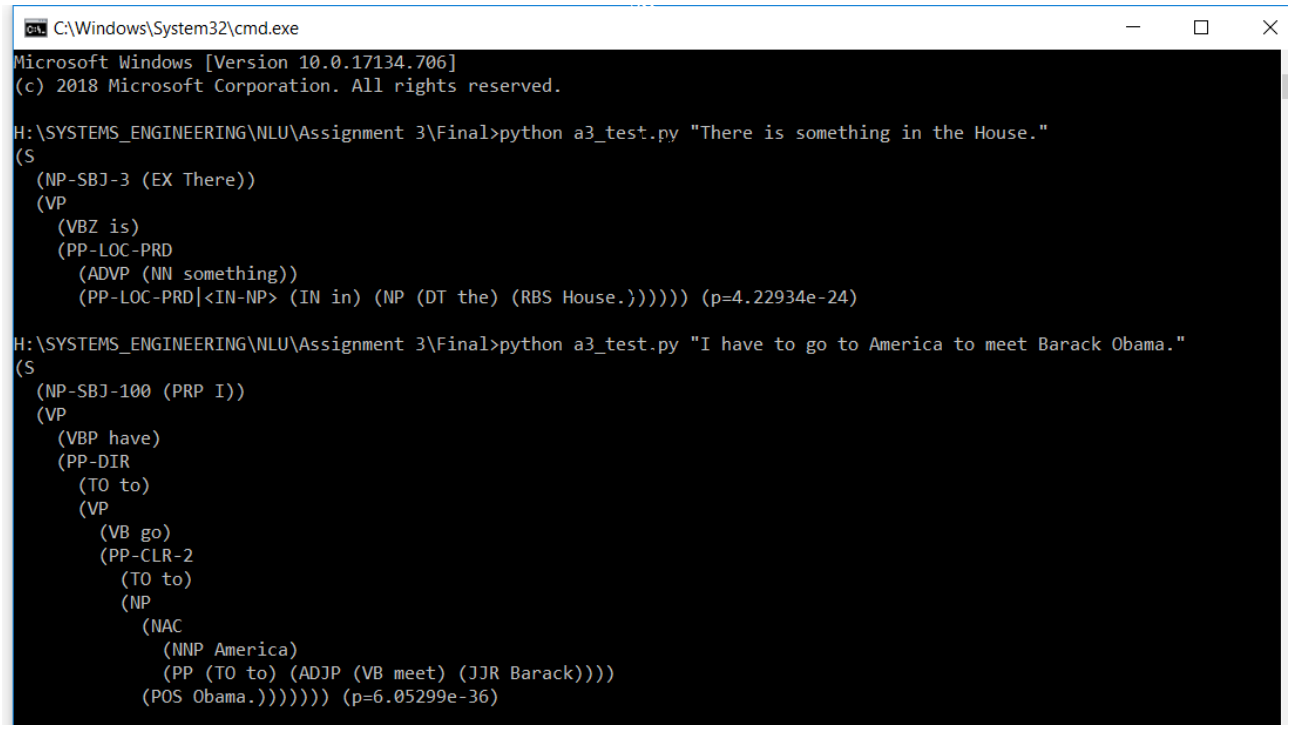
Table 1: **Task1 :Metrics calculated on the Test dataset**

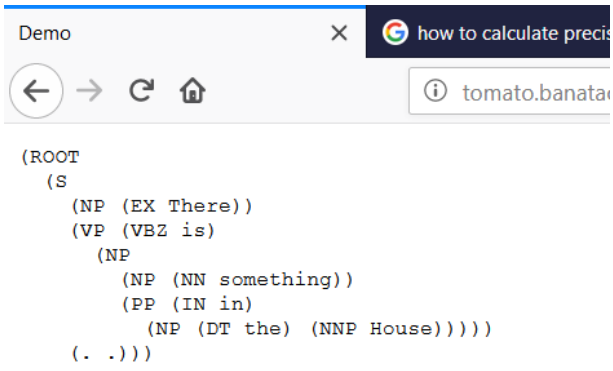
4 Probabilistic CYK Parser

CYK (CockeYoungerKasami algorithm) parses strings for Context free Grammars in bottom up fashion. It evaluates most probable parse tree.

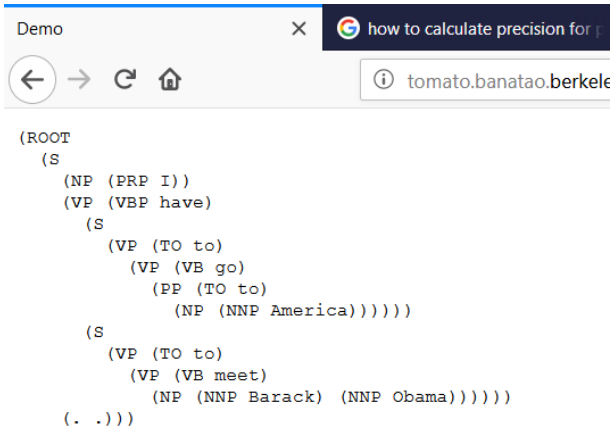
As the length of input increases, number of possible parse trees also increase. But not all of these trees are considered equally significant. CYK parser find the best, that is, most probable parse tree, or the first few best trees.

5 Task2 Results :





Above is the Parsing of the sentence: **There is something in the House.**



Above is the Parsing of the sentence: **I have to go to America to meet Barack Obama.**

6 Conclusion

It was observed that in sentence 2, ”Barack” and ”Obama” were not given correct Parsing Tags. I have used **Add 1 smoothing** in this experiment. I have also tried to change some rules to get the correct parsed outputs but Since, the Parser in this experiment are trained on very less data, it was not possible to get correct tags for this sentence.

References

1. Stanford Lectures on Parsing