Golas, Ekal

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

PROJECT PROPOSAL

nATURAL LANGUAGE PROCESSING



Table of Contents

[Summary 2](#_Toc444598638)

[Implementation Details 3](#_Toc444598639)

[Data source 3](#_Toc444598640)

[Input 3](#_Toc444598641)

[Expected Results 3](#_Toc444598642)

[Other details 3](#_Toc444598643)

# Summary

The following is proposed as the project idea:-

The idea is to implement a solution for statistical parsing. However:-

1. Statistical parsing uses a probabilistic model of syntax in order to assign probabilities to each parse tree.
2. Provides principled approach to resolving syntactic ambiguity.
3. Allows supervised learning of parsers from tree-banks of parse trees provided by human linguists.
4. Also allows unsupervised learning of parsers from unannotated text, but the accuracy of such parsers has been limited.

Hence the idea of Probabilistic Context Free Grammar (PCFG) is considered because:-

1. A PCFG is a probabilistic version of a CFG where each production has a probability.
2. Probabilities of all productions rewriting a given non-terminal must add to 1, defining a distribution for each non-terminal.
3. Resolves ambiguity by picking most probable parse tree.

Hence, the idea is: Given a probabilistic context-free grammar (PCFG) trained from a corpus, implement the CKY parser. The following would need to be taken care of:-

1. CKY must be modified for PCFG parsing
2. Must retain the most probable derivation of each constituent
3. Must set production probabilities to preserve the probability of derivations

# Implementation Details

## Data source

The following sources would be considered for obtaining the data required to implement this project:-

1. Penn Treebank – For the corpus
2. A PCFG file – Either a file that is trained from the Penn Treebank data or a resource available from the internet that can provide some relevant grammar rules
3. Class notes for the design of CKY algorithm

## Input

The following would be the input to the program:-

1. PCFG file that contains the grammar for parsing
2. A sentence as String

## Expected Results

The following would be expected assuming the input is valid:-

1. Generate the parse tree for the given sentence
2. (Optional) If time permits – also use the CKY parser as a recognizer which returns true if the sentence argument is in the language specified by the grammar, and False otherwise.

## Other details

Language chosen – Java 8

Expected time for implementation – 20 hours