NLP Project Outline

Mukund Choudhary - 2018114015, Tanish Lad - 2018114005

1. Problem Statement and Aim

Create a Dependency Parser for Marathi

The aim is to create a ML algorithm based dependency parser for Marathi i.e. based on Logistic Regression/ SVM etc. That takes in Chunked Input and finds out how they are related, what direction and tag should the relation get etc. It should also be able to make use of features like POS tags, Affixes (Marathi has a regular paradigm for most of its morphology), Word Embeddings etc.

2. Steps to Solve the Problem

We are still reading up all the past work on the problem and studying different algorithms but a rough sketch of how it is generally solved is as follows -

a. Get chunked data
    i. In our case, we found SSF chunked Marathi data from LTRC website
b. Use arc standard algorithm for parsing the dependency trees
c. Train the guiding ML algorithm (by guiding we mean that it guides about which action should be taken, shift or right-arc or left-arc or reduce or swap etc.)
    i. Which basically takes SSF as gold standard in our case
d. Finding out which features approximate to what and which ones to tune up/down

3. An outline of the algorithms, papers cited etc.

We are looking at the following papers, tutorials, algorithms etc.

a. Introductory - *NLP Programming Tutorial 12 -Dependency Parsing, Graham Neubig ; Transition-based dependency parsing, Sara Stymne*
    i. To gain basic understanding of different existing models/ model frameworks
    ii. To get a grip on what all would be the requirements, what are the common points between designs
    iii. To roughly plan for choosing one over the other when we decide for Marathi
b. Models - *Statistical dependency analysis with support vector machines, Yuji Matsumoto ; Universal Dependency Parsing from Scratch, Manning ; Transition-based Dependency Parsing with Rich Non-local Features, Nivre ; MaltParser: A Data-Driven Parser-Generator for Dependency Parsing, Nivre and Hall*
c. Theory and Comparative Study - *Algorithms for Deterministic Incremental Dependency Parsing, Nivre*
d. Reading (An)notations - *Stanford typed dependencies manual, Manning ; AnnCorra, LTRC*

4. What all we plan to do by the end of the semester.

*We plan to create a dependency pa*rser for the Marathi Language and match the current baseline to the extent possible.

- We first plan to create a basic working model according to the algorithms described in the above mentioned papers and get decent accuracies compared to hand labelled SSF data.
- As an improvement, we could then tune the model further, as now we are assuming it won't be perfectly language independent. Which would reflect the fact that this parser is for Marathi and features specific to it.
- If time persists, we plan to post-process the model and add some rules so that it becomes a Rule-based ML Hybrid Model.