SEMANTIC ROLE LABELING

PROJECT OUTLINE DOCUMENT

Submitted by:
Sumanth Balaji (2018114002)
Tanvi Kamble (2018114004)

I. Problem Statement

A SRL system has to label the arguments for each predicate of a sentence automatically.
The idea is to be able to implement a model (statistical or neural) that analyzes a sentence based on the role of the individual participants, from a combination of the part of speech, the dependency label and other linguistic information useful in determining theta role and associated information.

II. Aim

The aim of this project is to develop a linguistically grounded semantic role labeler based on the Hindi Dependency Treebank by implementing the paper 'Towards Building Semantic Role Labeler for Indian Languages' by Maaz Anwar Nomani and Dipti Misra Sharma

III. Steps to solve the problem

Step 1: Argument Identification - use a binary classifier ('argument' and 'not an argument') to identify a constituent in a dependency tree which represents the argument for the predicate in the argument predicate structure in a Hindi WordBank sentence.

Step 2: Argument Classification - Run a multi-category classifier to classify

the constituents that are labeled as arguments into one of the classes. The classifier is trained on multiple semantic labels present in the training data. We use Automatic Parses from Hindi parsers as input in this task with the features that it produces.

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

Towards Building Semantic Role Labeler for Indian Languages

Authors: Maaz Anwar Noman and Dipti Misra Sharma

The paper introduces a statistical semantic role labeler for Hindi and Urdu, two major Indian languages. Semantic Parsing is essentially the research investigation of identifying WHO did WHAT to WHOM, WHERE, HOW, WHY and WHEN etc. in a sentence and adding a layer of semantic annotation in a sentence produces such a structure. Different linguistic features of the data are tried and tested . Data is taken from Hindi TreeBank.

Two approaches were carried out and compared:

1. The two phase approach where, for a set of linguistics features marked on the input data, the first phase is identifying the arguments to a predicate in a sentence using a binary classifier and the second phase is essentially assigning a category to the identified arguments using multi-class classifier. The linguistic features include predicate, head, head-POS, phrase type, named entities and dependency.
2. This model directly classifies the arguments without identifying them. It was found that due to the overwhelming number of non-arguments present in the corpus, the classifier failed to classify the arguments correctly.

It was also observed that the hindi and urdu case markers provide important information about the token preceding it and the nature of the chunk/phrase in which they are present.

V. What all you plan to do by the end of the semester.

Our objective is to implement the first approach suggested by the paper (Towards Building Semantic Role Labeler for Indian Languages) and test the algorithms as specified which entails testing the 2 stage Semantic Role Labelling algorithm while taking into consideration different sets of features for the input data along with the baseline features that are considered and comparing the performance so produced with respect to both the stages. We also do extensive error analysis on the performances that are produced to draw all possible logical conclusions.