Dependency Parser for Hindi-Urdu Language

Shweta Singhal and Nishant Agarwal

INTRODUCTION

Dependency grammar (DG) is a class of modern syntactic theories that are all based on the dependency relation. Dependencies can be represented in various forms but the most common is the tree structure, where the words in a sentence are linked with also a representation of the POS tags. The amount of literature and work available online to read are solving problems in context to English Language. Over the past few years some amount of work has gone into solving problems related to other languages as well, but the Indic/Indian languages have been neglected for long. While discussing on different projects with Professor Vivek, his suggestion on building a dependency parser for Hindi-Urdu language was very interesting and exciting. Since there is no parser for these languages currently, It would be very exciting to work on the problem aiming to create an application.

RELATED WORK

Since parsers for English Language work very well and there has been some work with other languages. The below papers provide a starting point for building a parser for Hindi-Urdu:

1. "Max-Margin Parsing" by Ben Taskar, Dan Klein, Michael Collins, Daphne Coller and Christopher Manning.
2. "Less Grammar, More Features" by David Hall, Greg Durrett and Dan Klein.

DATA

We have acquired the dataset from "The Hindi-Urdu Treebank Project", a collaborative work by University of Colorado, IIIT Hyderabad and others. The dataset contains Hindi-Urdu data for development, training and testing for Interchunk and Intrachunk. The dataset is further divided into two formats CoNLL and SSF (Shakti Standard Format).

The CoNLL format is in Morphological format with the ID, WORD, POS tag, case, gender etc. provided. For further information one can refer the following website: http://universaldependencies.org/format.html

The SSF format is an XML based representation for each individual sentence with attributes present in each word node providing details about parent etc.

APPROACH

We aim to implement a base system using the approaches present in the "Max-Margin Parsing" and "Less Grammar, More Features" papers. The decision was taken as there is no parser for the Hindi-Urdu Language, so we would like to try the approach specified in the papers, as they give a very generalized approach to parsing for multiple languages and are very close or have beat the state of the art systems for those languages.

The approach described in the papers is the use of conditional random fields (CRF) over possible dependency trees for a given sentence S, where the features are the rules used in the tree. The start, stop and split indexes where rules are anchored is referred to as span(r), is used as surface features.

EVALUATION

For the evaluation of our system we would be using EVALB (http://nlp.cs.nyu.edu/evalb/). This software has been built on the paper written by Michael Collins and Satoshi Sekine. It is the standard method for measuring performance of parsing called PARSEVAL. The performance would be measured and represented using F-score, Recall and Accuracies of the system.