11-712: NLP Lab Report

Swabha Swayamdipta

April 25, 2014

Abstract

1 Basic Information about Hindi

The goal of this project is to build a dependency parser for Hindi. Hindi is one of the official languages of India, spoken widely in northern India. It uses a Devanagiri script. In some literature, it is also refered to as Modern Standard Hindi. It follows the subject-object-verb order.

2 Past Work on the Syntax of Hindi

Still to find a paper that discusses Hindi syntax, most papers describe experiments on various parsers.

2.1 Malt parser work at IIIT Hyderabad

2.2 CCG for Hindi Dependency Parsing

2.3 ICON 2010 - DeSR by Pisa

2.4 Self-training and co-training

3 Available Resources

Hindi treebank provided by IIIT Hyderabad (3000 sentences, parsed) - used for a reference to understand dependency grammar in Hindi.

WMT 2014 shared task data (about 1 million sentences) - unannotated corpora - can be used for building training data, and as test corpora.