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.

Panini model for IL is a dependency model

Karaka relations

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. Begum et al. (2008) DBL (2008)

4 Survey of Phenomena in Hindi

Dependency parsing for Hindi was a task at ICON 2009 and 2010.

Used Sanchay API tool for annotation

Made everything projective.

Rules:

PSP (ki) is the child of immediately preceeding/following subtree root

NST (baad) to immediately preceeding subtree

RP(bhi) to immediately preceeding tree
anything following PSP(ki) might be the root of preceeding subtree
brackets point to root of whatever is inside
VAUX to VM everytime
2 word expressions, first word and hyphen to second word
SYM to head of the rest of the tree
NN is the child of VM
QC is the child of immediately following NN
PRP is the child of NN
NN goes to NN (kisi tarah ki ghatna)
VM is the child of following NN (ghatna na ho iska prayas)
first name is the child of the second name. second name stands for the person.
CC is the parent for all around it
CC(ki) vs VM(kee)
INTF is the child of whatever comes next to it
hyphenated expressions : parent is whatever is on the right
2 similar expressions without comma (7 lakh 20 hazaar) - hazaar is the parent???
QF is the child of VM
DEM(Iss) is the child of NN
NEG is the child of VM
NST(paas) is the parent of nearby noun phrase
PRP(koi) is the child of nearby noun phrase
NN PSP NN (a ka b) - b is the parent
confusing case, bachhe hon ya vayask
if A then B, A-¿if-¿then¡-B
Amount of english words in the corpus is startling. Part of speech tags are wrong!

5 Initial Design

Semi-supervised dependency parsing (Collins).

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

Preprocessing involving removing of tokens, like punctuation.

References

Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008, 2008. The Association for Computer Linguistics.

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. Dependency annotation scheme for indian languages. In *IJCNLP* DBL (2008), pages 721–726.