Dependency Parsing - Hindi

Ayan Biswas - 2019121009
Veeral Agarwal - 2019114009
Mentor - Saujas Vaduguru
Course instructor - Prof. Manish Shrivastava

Interim submission:
Presentation video here.

Interim Report

Dependency parsing - Hindi

In this project, we are creating a dependency parser (hindi) for which the training dataset will be provided. A dependency parser identifies the syntactic dependency between words in a sentence. While a lot of literature exists in the dependency parser community in NLP, there are few well developed tools for indian languages.

Dataset

- Universal Dependencies Hindi Tree Bank
- Update: Using HDTB from LTRC instead of the Universal Dependencies version.

Interim Timeline

- Interim: Setup data preprocessing and Feature-Extraction pipeline. Implement a parser with greedy choice as a swappable module.
- Final: After testing various approaches, arrive at a robust model for Dependency Parsing in Hindi.

Work done

- Collected data from LTRC (interchunk->SSF->utf->news_articles_and_heritage) .
- Some manual cleaning and merged them in a single file.
- Dependency Parse Extraction
  - Simplifying data
  - Head Extraction from Chunks
  - Generating Parse from Head Extraction data

Initial data

1 ((NP <fs name='NP' drel='nmod:NP2'>

1.1 यहाँ PRP <fs af='यहाँ,pn,,,,o,,' name='यहाँ' posn='10'>

1.2 से PSP <fs af='से,psp,,,,,,' name='से' posn='20'>

))

2 ((NP <fs name='NP2' drel='jjmod:JJP'>

2.1 5 QC <fs af='5,num,any,any,,any,,' name='5' posn='30'>

2.2 किमी NN <fs af='किमी,n,m,sg,3,d,0,0' name='किमी' posn='40'>

2.3 दूरी NN <fs af='दूरी,n,f,sg,3,o,0,0' name='दूरी' posn='50'>

2.4 पर PSP <fs af='पर,psp,,,,,,' name='पर' posn='60'>

))

3 ((JJP <fs name='JJP' drel='nmod:NP3'>

3.1 स्थित JJ <fs af='स्थित,adj,any,any,,d,,' name='स्थित' posn='70'>

))

4 ((NP <fs name='NP3' drel='k1:VGF'>

4.1 वासुकि NNPC <fs af='वासुकि,n,m,sg,3,d,0,0' name='वासुकि' posn='80'>

4.2 ताल NNP <fs af='ताल,n,m,sg,3,d,0,0' name='ताल' posn='90'>

Simplified data

H NP NP nmod NP2

T यहाँ PRP यहाँ

T से PSP से

H NP NP2 jjmod JJP

T 5 QC 5

T किमी NN किमी

T दूरी NN दूरी

T पर PSP पर

H JJP JJP nmod NP3

T स्थित JJ स्थित

H NP NP3 k1 VGF

T वासुकि NNPC वासुकि

Head extracted

H यहाँ यहाँ NP PRP NP nmod NP2

H दूरी दूरी NP NN NP2 jjmod JJP

H स्थित स्थित JJP JJ JJP nmod NP3

H ताल ताल NP NNP NP3 k1 VGF

H अपने अपना NP PRP NP4 r6 NP5

H जल जल NP NN NP5 ccof CCP

Dependency extracted

H यहाँ यहाँ NP PRP NP nmod NP2 ; H दूरी दूरी NP NN NP2 jjmod JJP ; R ; nmod

H दूरी दूरी NP NN NP2 jjmod JJP ; H स्थित स्थित JJP JJ JJP nmod NP3 ; R ; jjmod

H स्थित स्थित JJP JJ JJP nmod NP3 ; H ताल ताल NP NNP NP3 k1 VGF ; R ; nmod

H ताल ताल NP NNP NP3 k1 VGF ; H है है VGF VM VGF NULL ROOT ; R ; k1

Work done against the timelines as mentioned in your project outline

- Target - Feature extraction and Parser Implementation
- Progress - Dataset processing pipeline - Formats, files, merges, simplification, head extraction, dependency extraction.
- Gap - Parser to be done during the next phase.

Plans till final submission

- Implement the parser.
- Train parser on SVMs or NN models using feature templates.
- Analyze performance of the parser on test data.

Thanks!