Morphological Analysis of Hindi Language

Group Members

1) Sanket Kale (22111052)
2) Saqeeb (22111053)
3) Abhinav Kuruma (22111401)

Introduction

- Morphology refers to formation of words by focusing on their internal structure.
- Morphological analysis is concerned with analyzing individual words into their components.
- Effective implementation of morphological analyzer can be seen in language which is rich in morphemes such as Hindi.
- We propose to create a morphological analyzer for Hindi.
- It finds out the pos tag of each word in sentence given a sentence and gender,case,lemma and number given a word.

Problem Statement

- To find out the pos tag,lemma and other morphological features of the word such as gender,case and number.

Proposal

- Finding the pos tags of each word given a sentence using HMM
- Finding the lemma given a word using rule based approach
- Finding the gender,case and number of given word using both rule based and statistical approach

Process Flow

1. **POS tagger using HMM**
   We created POS tagger for given sentence using HMM algorithm.
2. **Lemmatizer for each word**
   We are finding out the lemma of given word for the hindi dataset.
3. **Gender,Case and Number prediction**
   We are using rule based approach and deep learning model for predicting these features of given word.

Data Collection

- We are using Hindi labelled dataset provided by AI4Bharat in conllu file.

LINK : `https://github.com/UniversalDependencies/UD_Hindi-HDTB/tree/master`

Lemmatization

- We created a lemmatizer for Hindi language using rule-based approach.
- If the given input word is not ending with any of the suffixes then we are assuming that it is a lemma trivially
- If input ends with a suffix then we are removing the suffix and checking if the remaining word is a root word.
- If that is not a root word then We add another suffix so that maybe after adding it becomes a root because for some examples we have to remove a suffix and add another suffix so that it becomes a meaningful word.

RULES

Lemma	Suffix
नज़र	ें
सड़क	ोंं
लड़की	-
खुश	ी
भारत	ीयता
मजदूर	ी
बालिका	ओं
विश्वास	नीय
सफल	ताओं
लड़का	ोंं
संशोध	न
तिजोरी	यों
लड़की	ियाँ
ज्यादा	-

		Rule application	
Word	**Root**	**Extraction of suffix**	**Addition of character**
लड़कियों	लड़की	ियों	ी
कहानियों	कहानी	ियों	ी
कवियों	कवि	ियों	ि (exception)
चिड़ियों	चिड़िया	ियों	िया (exception)

POS Tagger

- We are using HMM model for predicting POS tags of all words given a sentence.
- HMM stands for Hidden Markov Model
- Experimentally it is proven that HMM gives much better accuracies than traditional approaches.
- We used viterbi algorithm which uses dynamic programming for speeding up HMM execution.
- We got around 92% accuracy for POS tagging using HMM for hindi corpus.

Gender,Case and Number Prediction

- We are using two approaches for finding out these features,which are rule based and deep learning model based.
- We created a dictionary of certain suffixes which are very common in one type of class like feminine gender or plural number.
- If given word’s suffix matches with that word we are predicting its respective class.

- In our second approach we created a deep learning model in which we created a sequential model containing one hidden layer having 128 nodes.
- Our deep learning based approach is outperforming rule based approach.
- We got around 78% and 83% accuracies for gender and number respectively using rule based approach.
- We got around 90%,92% and 65% accuracies for gender,number and case respectively using deep learning based approach.

Results

	ACCURACY using rule-based	ACCURACY using Deep learning/HMM
GENDER	78.52%	90.5%
NUMBER	82.97%	92%
CASE	NA	65.37%
POS TAGGING	NA	92.04%
LEMMATIZATION	44.9%	NA

Table: Accuracies of different attributes by different methods

Conclusion

- We proposed morphological analysis on Hindi dataset by providing pos tags,lemmas,gender,case and numbers of each input hindi word.
- We used both deep learning and rule based approaches for some features and we compared the results.