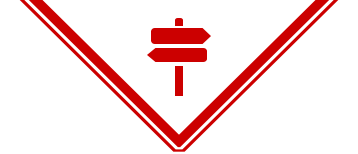
Hindi Lemmatizer

By:

Pranjal Kumar Srivastava (22111046)
Mayank Devnani (22111042)
Pradeep Chalotra (22111045)

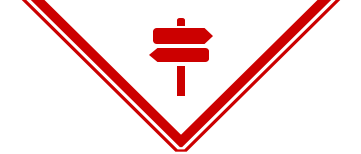
Presented to :

Dr. Arnab Bhattacharya



LEMMATIZATION

- Lemmatization means mapping a surface word in a context to its root form
- Mandatory pre-processing module where semantic processing is needed
- Different from stemming



RELATED WORK

- **Publicly Available Dataset**
  - UD treebank for Hindi created at IIIT Hyderabad, India containing the mappings between words and their lemmas
- **Lemmatization**
  - Rule based approach is one of the most accepted lemmatizing algorithms
  - Clustering based approach used for discovering the equivalent classes of root words
  - Work done using deep neural network architecture

HINDI LEMMATIZATION

- The lemmatizer that we built was using *rule based approach*
- Approach Used
  - We have used rule based approach for extracting the suffixes.
  - In rule based approach we have created many different rules for striping/appending of suffixes .
- Suffix Generation
  - For suffix generation we are going through the dataset and examined the changes after removing the suffix.
  - This led to the making of rules
  - The word 'लड़कियों' is derived by adding the suffix 'ियों'
  - Similarly there are many others words with same suffix

HINDI LEMMATIZATION

- Rule Generation
  - After the generation of suffix we have developed rules
  - We developed rule in such a way that we removed the suffix from the word and if required addition of ‘maatra’ or character take place
  - Example: लड़कियों - ियों + ी= लड़की
- Algorithm steps
  - Check the word in database
  - If present then display it
  - Otherwise apply the rules
    - Strip the suffix and if required then add suffix

HINDI LEMMATIZATION

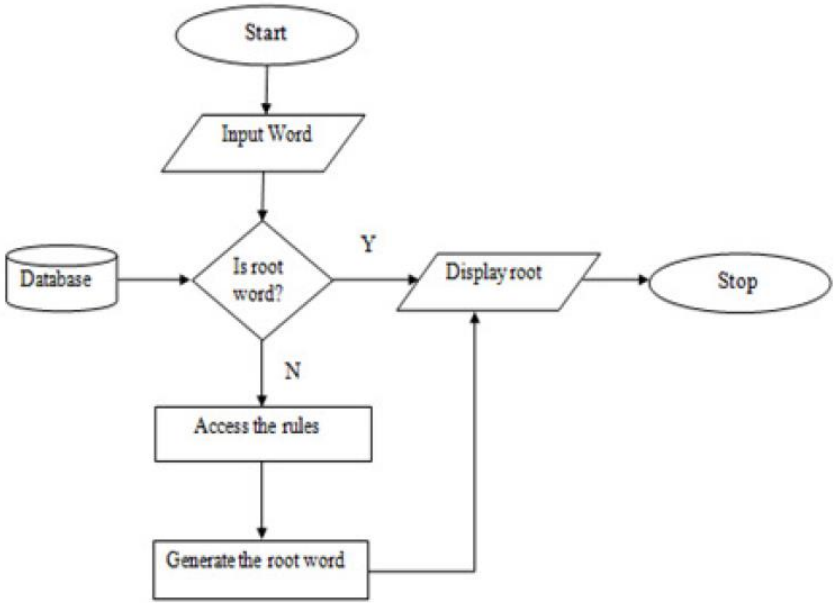


Figure: Flowchart for Rule Based Lemmatizer

DATASET GENERATION

FastText Vector Embedding

- Extract the list of lemma given in the UD treebank
- Find the FastText vector embeddings for these lemmas
- Use these embeddings to find the 5 nearest words to these lemmas
- Pick only those words for whom the complete lemma exists in their prefix. These are valid words
- Create a trigram of words from a Hindi corpus
- Each valid word is searched in this trigram to find its context
- The context for a word is its immediate neighbour words present in trigram
- Generated 9,345 word-lemma mappings with context

DATASET GENERATION

- **From UD Treebank**
  - Parsed UD treebank dataset
  - Generated word-lemma mappings with immediate neighboring words as context
  - Generated 17,451 word-lemma mappings this way

- Generated total 27,010 word-lemma mappings with context for Hindi language in Devanagari script

DATASET GENERATION EXAMPLES

S. No	DERIVED WORD	CONTEXT	LEMMA
1.	टेलीफिल्मों	कई टेलीफिल्मों में	टेलीफिल्म
2.	मनपसंदीदा	अपना मनपसंदीदा आइलाइनर	मनपसंद
3.	मच्छरो	रस मच्छरो को	मच्छर
4.	पर्यटनीय	उस पर्यटनीय अनुभव	पर्यटन
5.	गतिविधिओ	आतंकी गतिविधिओ और	गतिविधि

Table: Word-Lemma Mappings with Context

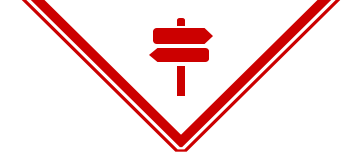
RESULTS

Hindi Lemmatizer

- Most of the rules violated both the exceptional and general rules.
- Accuracy of the system was calculated using the equation
- Accuracy =(number of correctly found lemma/total no of words)*100
- Accuracy obtained is 58.98 %

Dataset Generation

- Total 27,010 word-lemma mappings with context for Hindi language in Devanagari script



FUTURE WORK

- If a bigger list of Hindi root words is present, more word-lemma mappings with context can be generated to create a bigger dataset
- Rule based lemmatizer developed here can act as a baseline using which more complex Deep Learning based techniques can be explored

Thank You!