ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

A FRAMEWORK PROPOSAL FOR HINDI LANGUAGE PROCESSING

Ashwani Gupta
Assistant Professor, CSIT Department,
MJP Rohilkhand University, Bareilly-UP (India)
ashwani.gupta@mjpru.ac.in

ABSTRACT

Most of the Indians use Hindi as their first language, so it is necessary to develop application for Hindi language. An attempt is present here for Hindi natural processing. Paper is comprised as follows, Firstly need for NLP for Hindi language is presented, A frame work for Hindi NLP is presented in section 2.A few major recourses for Hindi corpora are presented in section 3, Section 4 presents various challenges during Hindi NLP, a conclusion is presented in last section.

NATURAL LANGUAGE PROCESSING FOR HINDI LANGUAGE

Most Indian are bilingual and study more than one language in the school. They have Hindi as their first language, followed by Marathi, Assamese, Telugu, and Tamil etc. Among the world's fast-growing economies and one with the second largest population, the Indian market is garnering considerable interest and is on the radar of internet and software companies. There is considerable attention around developing utility applications that rely on the understanding of language to function as bots in call canters, customer services, search, virtual agents etc. across multiple channels including voice, web and social. It is comparatively easy for computers to process the data represented in English language through standard ASCII codes than other Indian languages. Such compatibility development is too tough for Indian languages. Much research is being carried out to facilitate users to work and interact with computers in Hindi and other languages [1].Natural Language Processing, usually shortened as NLP, is an interdisciplinary field that deals with the interaction between computers and humans using the natural language [4].NLP is a way for computers to analyze, understand and derive meaning from human language in a smart and useful way. So there is a need to apply NLP for Hindi and other languages.

A FRAMEWORK FOR HINDI NLP

A flowchart for natural language processing is shown by the following flowchart. It consists of various stages i.e. tokenization, Transliteration, pre-processing, analysis either by data-driven approach or rule-based approach. Any task viz Machine Translation, Automatic Summarization; Question-Answering can be performing after above analysis. The raw data is supplied in the form of input text or available text. This input data is pre-processed by pre-processing technique(s). Rule-based analysis approach or data-driven analysis approach is used for performing analysis and to develop application.

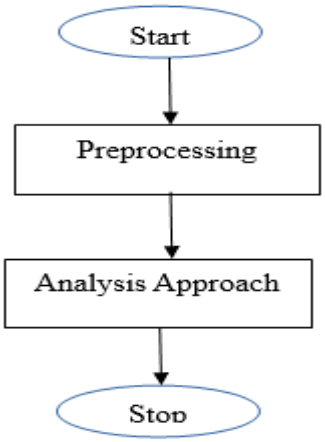

Figure 01- Framework for Hindi NLP

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

Pre-processing
Pre-processing is important step in Hindi natural Language Processing. It is used to convert the raw data in a form from original data to a format, that can be easily and efficiently used in further processing steps. Following are some pre-processing techniques for Hindi Language dataset:

Tokenization
Tokenization is the first step of Hindi NLP. It is the mechanism of fragmenting or splitting the data into smallest basic unit which need not be decomposed later processing. Tokenization is the mechanism of splitting or fragmenting the sentences and words to its possible smallest morpheme called as token. Morpheme is smallest possible word after which it cannot be broken further. Tokenization can be performed at three level either word-level, sentence-level or n-gram level as per the problem [7].N-word tokenization depends on number of words taken together, if n = 1, unit is called unigram, if n = 1, unit is called bigram, becomes n-gram for n > 2. In NLP fir Hindi studies, it is conventional to concentrate on pure analysis or generation while taking the basic units, namely words, for granted.

Segmentation
Text segmentation is the process of dividing written hindi text into meaningful units, such as words, sentences, or topics[12]. The term applies both to mental processes used by humans when reading Hindi text, and to artificial processes implemented in computers, which are the subject of natural language processing. In Hindi language processing Sentence is segmented by identifying the boundary of sentence which ends with 'I' end marker.

Stop-word Removal
Frequency of stop-words, occured much in a document. These words have no relevant meaning but some-time creates a higher frequency in the frequency list. Remove of the stop words reduces the document size to a considerable extent and saves time in text processing in Natural Language. Some of the stop-words for Hindi language are as follows:
"और" (and), "है" (is), "के" (of), and "में" (in).

Stemming
Stemming is performed to obtain the stem or radix of those words which are not found in dictionary. If stemmed word is present in dictionary, then thata genuine word, otherwise it may be proper name inflected word or some invalid word [10]. The main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. Stemming is widely uses in Information Retrieval system and reduces the size of index files. We can say that the goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Lemmatization
Stemming is the process of clipping off the affixes from the input word to obtain the respective root word, but it is not necessary that stemming provide us the genuine and meaningful root word. Lemmatization It is similar to the stemming technique but the word is reduced to an acceptable form in the language after the removal of suffixes and pre-fixes. The reduced word is called "Lemma"; it is valid and accepted by the language [11].

POS Tagging
Code-mixing occurs when a person changes language (alternates or switches code) below clause level, so internally inside a sentence or an utterance. This phenomenon is more abundant in more informal settings — such as in conversational spoken language and in social media text — and of course also more common in areas of the world where people are naturally bi- or multilingual, that is, in regions where languages change over short geospatial distances and people generally have at least a basic knowledge of the neighbouring languages. As its name suggests it stands for identification of Part of Speech like; noun, pronoun, verb, etc.POS Tagging is very useful in translation into some other language [9].

Unicode Normalization
Unicode is the Universal character encoding standard, which represents the text characters by a unique hexadecimal value. Characters by a unique hexadecimal value. This representation is used for information

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

processing. Some sequence of Unicode characters are equivalent to single abstract Unicode character, this multiple representations for an abstract character leads to complication. To eliminate these non-essential differences Unicode normalization is performed during pre-processing specially in Hindi language[5].

Analysis Approach
Broadly any one of the following two approaches can be used for analysis:

Rule-based Approach
Rule-based Approach is an early approach used before data-driven This approach exists before the data-driven approach. The lexical and morphological analyses using techniques like regular expressions, Suffix striping and so on are applied after pre-processing. In rule based natural language processing approach, the set of rules and patterns guide the machine to translate the language. E.g.: Hindi has the language structure of SOV (Subject,object,Verb) [6].Rule Based Machine Translation (RBMT) technique translates source language to target language using various set of rules and bilingual dictionary in machine translation task of language processing.

Data-driven Approach
Data-driven Approach derives information from representative-examples and apply neural networks or statistical tools. Statistical Tools Based Approach The data is analyzed by some statistical methods and some statistical parameters are computed i.e. distance metric, probability-metric [3] etc.

Neural Network Based Approach It is an emerging method because suggests better results in language processing tasks. This approach uses artificial neural networks architecture i.e. Recurrent Neural Network (RNN) [8], Long-Short Term memory (LSTM) [2] etc.

Applications
These approaches can be applied to perform i.e Sentiment analysis(SA),Name Entity Recognition(NER), Machine Translation(MT), Summerization, Question-Answering System(QA) System etc.

Sentiment Analysis
Finding the contextual polarity of people's sentiment's, opinion, attitudes, appraisal and emotions in text is the task of Sentiment Analysis (SA). It gives polarity of text in terms of Positive, Negative and Neutral. Mainly researchers have studied sentiment analysis at different levels as Document level, Sentence level and Aspect level sentiment classification [13]. In Document level sentiment classification is done on whole documents which results in document may be Positive, Negative or Neutral. In Sentence level classification individual sentences indicates polarity as Positive, Negative or Neutral. And Aspect level classification considers aspect terms present in sentences which results in opinion related to aspect terms, Aspect level sentiment analysis is divided into different task as extraction of aspect, extraction of entity and sentiment classification of aspect.

Name Entity Recognition
The objective of NER is to identify and classify every word/term in a document into some predefined categories like person name, location name, organization name, miscellaneous name (date, time, percentage and monetary expressions etc.) and "none-of- the-above"[14]. The challenge in detection of named entities (NEs) is that such expressions are hard to analyze using rule-based NLP because they belong to the open class of expressions, i.e., there is an infinite variety and new expressions are constantly being invented.

Machine-Translation
Like translation done by human, MT does not simply substituting words but the application of complex linguistic knowledge; morphology, grammar, meaning all this things is taken into consideration. Generally, MT is classified into various categories: Direct based, rule-based, corpus based, statistical-based, hybrid-based, example-based, knowledge-based, principle-based, and online interactive based methods [15]. At present, most of the MT related research is based on Rule based approaches because rule based is always extensible and maintainable.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

Summerization
Text summarization is defined as a task of minimizing a text that is produced from one or more texts such that the actual significant information in the texts is not lost. A text summarization tool compresses the text and displays only the important content to the user[16]. Using text summarization, decisions can be made in lesser time and the core of the document be understood.

Question-Answering System
Automatic text summarization is the process of shortening a given text, by a computer program, without losing the actual information to be conveyed. There are two widely used methods in text summarization—Extractive and Abstractive. Extractive summarization extracts the texts and creates the summaries by reusing portions (words, sentences, etc.) of the input text[17], while abstractive summarization is those which create the summaries by re-generating the significant content of theinput text. In case of summarization system, there are summarises from singledocument or multiple documents and these kind of summarization system is calledas multi document summarization system.

RESOURCES
Following are resources found for Hindi Language:-
1. Centre for Indian Language Technology, IT Bombay http://www.cfilt.iitb.ac.in/Resources.html
2. Dataset Source-EMILLE https://www.lancaster.ac.uk/fass/projects/corpus/emille/
3. Speech Dataset in Hindi Language https://ieee-dataport.org/open-access/speech-dataset-hindi-language-0
4. Kaggle-HindiEnglish Corpora https://www.kaggle.com/aiswaryaramachandran/hindienglish-corpora

CHALLENGES
There are several challenges faced in all the stages of the anguage processing tasks because of differences in grammar and phonetics in hindi. The challenges faced in hindi language processing are as follows:
1. Language structure i.e. order of the words in the sentences will differ from one language to another E.g.: Subject Verb Object (SVO) (English) and SOV (Hindi).
2. Ambiguity in translation or transliteration of regional language words.E.g.: In English-Hindi translation, the word mount translates but Everest remains same. In English-Kannada both the words are just transliterated.
3. Grammatical variations between languages lead to ambiguity.
4. Judging of speakers intention is difficult. Meanings of sentences or words vary with the speakers intention (like sarcasm, sentiment, metaphor, etc.).
5. Code-Mixed language processing is challenging as user uses multiple languages in a sentence or an utterance.e.g. Main kal movie dekhne jaa rahi thi and raaste me I met Sudha.

CONCLUSION
An approach is briefly presented to develop applications based on Hindi language processing. Various components are briefly described to develop , all types of pre-processing may not be applicable to input data, depend upon data type or output produces after pre processing. Either data-driven approach or rule-based approach can be applied before the application developement.Above framework would be beneficial for various developer and to find approaches at different module development.

REFERENCES
1. S Amarappa and SV Sathyanarayana. Kannada named entity recognition and classification (nerc) based on multinomial na\" ive bayes (mnb) classifier. arXiv preprint arXiv:1509.04385, 2015.
2. Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. Emotions are universal: Learning sentiment based representations of resource-poor languages using siamese networks. ArXiv preprint arXiv:1804.00805, 2018.5
3. Rakhi Joon and Archana Singhal. Analysis of mwes in hindi text using nltk. International Journal on Natural Language Computing, 6:13–22, 02 2017.
4. Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. Journal of the AmericanMedical Informatics Association, 18(5):544–551, 2011.
5. Pramod Pandey. Akshara-to-sound rules for hindi. Writing Systems Research, 6(1):54–72, 2014.
6. Raj Nath Patel, Prakash B. Pimpale, and M. Sasikumar. Machine translation in indian languages: Challenges and resolution. Journal of Intelligent Systems, 28(3):437 – 445, 01 Jul. 2019.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

7. Snigdha Paul, Nisheeth Joshi, and Iti Mathur. Development of a hindi lemmatizer. arXiv preprint arXiv:1305.6211, 2013.
8. S. Seshadri, A.K.a b Madasamy, S.K.a b Padannayil, and M. A. Kumar. Analyzing sentiment in indian languages micro text using recurrent neural network. 2016.
9. Vijay Kumar Sharma and Namita Mittal. Exploring bilingual word vectors for hindi-english cross-language information retrieval. In Proceedings of the International Conference on Informatics and Analytics, ICIA-16, New York, NY, USA, 2016. Association for Computing Machinery.
10. Satyendr Singh and Tanveer J Siddiqui. Sense annotated hindi corpus. In 2016 International Conference on Asian Language Processing (IALP), pages 22–25. IEEE, 2016.
11. Jonathan J Webster and Chunyu Kit. Tokenization as the initial phase in nlp. In COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics, 1992.
12. YLehal, Gurpreet Singh. "A word segmentation system for handling space omission problem in urdu script." Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing. 2010.
13. Arora, Piyush. "Sentiment analysis for Hindi language." MS by Research in Computer Science (2013).
14. Jain, Arti, et al. "Research trends for named entity recognition in hindi language." Data Visualization and Knowledge Engineering. Springer, Cham, 2020. 223-248.
15. Kumar, Pankaj, Sheetal Srivastava, and Monica Joshi. "Syntax directed translator for English to Hindi language." 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). IEEGupta, Manisha, and Naresh Kumar Garg. "Text summarization of Hindi documents using rule based approach." 2016 international conference on micro-electronics and telecommunication engineering (ICMETE). IEEE, 2016. E, 2015.
16. Gupta, Manisha, and Naresh Kumar Garg. "Text summarization of Hindi documents using rule based approach." 2016 international conference on micro-electronics and telecommunication engineering (ICMETE). IEEE, 2016.
17. Sinha, R. Mahesh K. "Automated mining of names using parallel Hindi-English corpus." Proceedings of the 7th Workshop on Asian Language Resources (ALR7). 2009.