POS Tagger for Hindi Language

Submitted By

Salman Khan

13MCEN13

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

POS Tagger for Hindi Language

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering (Networking Technologies)

Submitted By

Salman Khan

13MCEN13

Guided By

Prof. Tarjni Vyas

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

Certificate

This is to certify that the major project entitled **"POS Tagger for Hindi Language"** submitted by **Salman Khan (Roll No: 13MCEN13)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering (Networking Technologies) of Institute of Technology, Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Tarjni Vyas
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof. Gaurang Raval
Associate Professor,
Coordinator M.Tech - CSE
Institute of Technology,
Nirma University, Ahmedabad

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr K Kotecha
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Salman Khan**, Roll. No. **13MCEN13**, give undertaking that the Major Project entitled "**POS Tagging for Hindi Language**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering (Networking Technologies)** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Prof. Tarjni Vyas

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to my guide **Prof. Tarjni Vyas**, Assistant Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for her valuable guidance and continual encouragement throughout this work. The appreciation and continual support she has imparted has been a great motivation to me in reaching a higher goal. Her guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr K Kotecha**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Salman Khan

13MCEN13

Abstract

Part-of-Speech tagging, also called as POS tagging, can be defined as method of tagging language words according to its grammar category. For English language, there are a lot of POS taggers available currently, but they cannot work as same for Hindi language as these two languages have so many differences such as their formation is totally different. Many have attempted to develop a good POS tagger for Hindi but the structure and its complexity makes it very difficult. Basic aim is to perform part of speech tagging for Hindi Language. One of the reasons behind this is that the numbers of Hindi files are increasing in bulk on World Wide Web daily. So, it has become essential to process these files. POS tagging of Hindi language is in itself a very difficult task because not much research has been done in this area. Building a Hindi POS tagger need a good amount of linguistic knowledge. Based on this, Hindi POS tagger has been built using various approaches. One of the approaches is to assign tag using machine learning algorithm. Our aim is to build POS tagger using one such approach. We are using simple Hidden Markov Model approach to achieve Hindi POS tagging.

Abbreviations

POST	Part-Of-Speech Tagger
HMM	Hidden Markov Model
CRF	Conditional Random Fields
IIIT	International Institute of Information Technology

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Natural Language Processing

Natural language processing (NLP) is a field of artificial intelligence, computer science, and linguistics which concerns about human and computer communication[1]. Normally NLP can be said to be the area of interaction between human languages and computers. There are many challenges in NLP mainly natural language understanding, i.e., Computer understanding of human languages, and others involving generating this language[1].

In short, NLP Interface provides humans to communicate with computers in languages like Hindi, English Gujarati, Tamil or any other language used between humans. One of the major goal today is to build a translation system between these languages. This is required because there is need of communication between people from different regions[2].

Today many NLP algorithms use machine learning algorithms[3], most used one is statistical machine learning technique. The pattern used by machine learning is far different against the most of the previous attempts for language processing. Previous implementations of this processing tasks involved the large amount of hand coded rules. The machine learning patterns uses automatic learning of such rules through analysis of data using real life examples instead of using such general learning algorithms. A corpus is collection of document files which are hand-written with proper values.

The standard NLP tasks are Part-of-speech tagging(POS), Chunking(CHUNK), Name entity recognition(NER) and Semantic role labelling(SRL). POS tagging includes syntactic labelling of words with unique tag.

1.2 Stemming

Stemming is defined as linguistic operation which trims related words into a single term without performing complete formation analysis. Stemming helps in Information retrieval systems for performance improvement. Plus, no. of terms also get reduced in IRS systems such that indexing becomes easier[4].

The word stem need not be a valid form of word, it is important to mark similar words at same stem[5]. Reducing similar words to same stem allows using one variant for indexing purpose. This improves recall. But sometimes it reduces performance of system[6]. Hindi is very difficult language as change in form of the word changes meaning in context. Hindi words have many number of forms. These forms are mostly formed appending suffix to base or stemmed word.

Previously English stemmer was made by Julie Beth Lovins in 1968. Then The Porter stemming algorithm was proposed by Martin Porter in 1980 and still it is the most widely used algorithm for stemming for English language. Both of these stemmers are rule-based and are used for less inflectional languages such as English but same cannot be used for free order languages like Hindi.

Since last two decades there is a lot of work done in the field of unsupervised learning of morphology. Goldsmith, in 2001, proposed an unsupervised approach to learn the morphology of language on the basis of Minimum Description Length (MDL) framework which aims at representing the data in very compact manner. But there is not good work been reported for stemming for Indian languages like Hindi as compared to English and other European languages. Ramanathan and Rao in 2003 used a hand crafted suffix list and performed longest match stripping to develop a Hindi stemmer. Majumder in 2007 developed YASS: Yet Another Suffix Stripper which performs a clustering-based approach based on string distance measures and more importantly, it does not requires linguistic knowledge. Pandey and Siddiqui proposed an unsupervised stemming algorithm for Hindi based on Goldsmith's approach in 2008[6].

Stemming is mostly used in IR systems. Lets take an example, suppose some user enters *stemming* as query word, it is understood that he/she expect retrieved documents to contain the words *stemmer* and *stemmed* also. This will improve recall as number of documents retrieved will be higher. Also, Many terms are mapped to one as a result of

stemming, so it decreases Information retrieval system's index file size.

Approach proposed by Frakes and BaezaYates in 1992 is storing all possible index terms and their relative stems in a table and then stem the terms with lookup through table. This improves accuracy and efficiency in terms of speed, there is much more difficulty to in getting such data, which is usually not available. Also this approach will keep stemmer to only the table and domain dependency is not considered. Many dynamic approaches use statistical measures. Other two approaches are successor variety stemmers and ngram stemmers. Successor variety stemmers were proposed by Hafer and Weiss in 1974 shows approach of the distribution of morphemes in a huge text data based Identification of morpheme boundaries. N-gram stemmers proposed by Adamson and Boreham in 1974 in which stemming is done on the basis of number of n-grams which are common in terms. Affix-removal stemmers are based on removing word suffixes plus prefixes. These stemmers iteratively remove largest sequence of words from a word in reference to a set of rules. Some affix-removal stemmers also able to transform the stem outcome in some cases such as proposed by Porter in 1980 and Paice in 1974 [4].

1.3 Part of Speech Tagging

Part of Speech (POS) Tagging is the first step in the development of any NLP Application. POS tagging which is also known as word category disambiguation or grammatical tagging, is a technique of appending part of speech category with each token or word which helps in deriving relationship between these words and finding out relationship between words in given paragraph or sentence[7]. POS tagging plays important role in language processing which helps algorithms to use different terms and its hidden part of speech shown by its tag.

Most words shown in content has got uncertainty connected with it regarding their part of speech. For example **bank** is both noun and verb. So this word will be tagged in both of this form. But it is required and must to know the context before deciding the proper Part of speech. This opens gate for need of more analysis.
Main categories in which Hindi words tagged are noun, adverb, verb and adjective. Hindi POS tagging takes words role individually and also in the sentence it resides.

Tagging accuracy[8] depends on four factors:

- size of training data
- tag set
- The difference between training corpus and dictionary to the corpus of application
- Unknown words.

POS tagging is a very much complicated task which has to deal with issues such as POS tags ambiguity and also need to handle "lexical absence" (proper nouns, spelling variations, foreign words, derived morphed words and many other unknown words) [9].

1.4 Hindi WordNet

Hindi WordNet, which is in other words a rich computational lexicon, is widely being used for many Hindi NLP applications. Hindi WordNet is considered as a machine readable dictionary which follows psycholinguistic principal unlike standard alphabetical dictionary which organizes vocabularies using morphological similarities[10]. Vocabulary information is defined as word meaning using structures in Wordnet. Also one can say that Hindi WordNet is a collection of different lexical and semantic relations between the Hindi words. Each word has a synonym set (synset), defining one category of vocabulary. Hindi WordNet contains words in form of synsets and they are useful when there lies ambiguity in words, that is, more than one meaning of single words resides in sentence or paragraph. Hindi WordNet mostly contains content words and open class category of words such as Noun, Verb, Adjective and Adverb.

In other words, WordNet is a large lexical database. It cantains Nouns, verbs, adjectives and adverbs being grouped into sets of cognitive synonyms also called synsets, a distinct concept is expressed by each one. Synsets are interlinked based on conceptual-semantic and lexical relations. This network of related words and their conceptual data can be seen using a browser. WordNet are freely and publicly available for different languages for download and use. WordNet's structure makes it a useful asset in field of computational linguistics and natural language processing. WordNet superficially resembles a thesaurus, which assign words to groups through their meanings.But there are some differences. Let us look at some. WordNet interlinks word formsstrings of letters and also specific senses

of words. This results in words that are found in close proximity with others in the network, are semantically disambiguated. Secondly WordNet labels the semantic relations in between words, whereas thesaurusthe groupings of words does not have any explicit pattern but of similarity[1]. The main relation existing in between words in WordNet is nothing but synonymy, like for example shut and close or car and automobile. Synonyms are the words gives similar concept and can be interchanged in many contexts, this are grouped into unordered sets called as synsets. English WordNet has 117 000 synsets which are linked to other synsets with help of conceptual relations. In other words, each synset contains a brief definition ("gloss") and in almost all cases gives one or more short sentences shows the use of the synset members. Word forms with many different meanings are represented in in many distinct synsets. This means the every form-meaning pair is unique inside WordNet[1].

1.5 Machine Translation - Parsing

Machine Aided Translation is a process of generation of parse trees with respect to rules given as a input to the parser[2].

Parsing checks for correctness of structure of input language. In order to do this, grammar rules are created or we can say that as aset of production rules where each one has a structure like[2]:

$$< N - T > \rightarrow < N - T >|' T' \tag{1.1}$$

In above N-T stands for non-terminal and T for terminal. If the input sentences is correct when checked against set of these grammar rules then sentence is syntactically correct. Given sentence is parsed using trees, called Parse trees. Machine Translation has been occasionally addressed in the literature in the Last few years; but no widely accepted solution seems to have emerged till date[11].

Machine Aided Translation includes generating correct parse trees for the language rules given as input to the parser. This is a challenge because the attachment of karaka, showing verb and noun group attachment, gives a major threat. Other challenge is the identification of Gender because Hindi contains only Masculine, ending with -a and Feminine, ending with -i. There is no third inanimate gender as such in English like itselt. Also many times the result has many parse trees for single sentence which shows ambigu-

ity. Other issues include ambiguity and multiple translation for one input sequence and cases where parser cannot identify complex structures of input sentences. The ambiguity in parsing can be solved by using contextual information of the sentence and analysing its semantics.

Some approaches are developed for machine translation which uses one or more of the following techniques[2]:

- Rule-Based Rules for parsing and translation are written in specific format, such as context free grammars.
- Example-Based This approach implies the principle of analogy in between two or more parallel corpora of different languages which are involved in the translation process. It can be source , destination or intermediate language form.
- Statistical Method This approach includes calculation of probabilities of words of source, getting translated to one or more related words in the target language. The word with the highest probability is selected as target. This method is applicable at the word-level as well as the sentence-level.

1.6 Hidden Markov Model(HMM)

A hidden Markov model (HMM) can be defined as statistical Markov model where the system being modeled is supposed to be a Markov process with unobserved states or hidden states. A HMM can also be looked as the simplest dynamic Bayesian network. The mathematics behind HMM was developed by L. E. Baum and coworkers[1].

A POS tagger based on HMM assigns the best tag to a word by calculating the forward and backward probabilities of tags along with the sequence provided as an input. Suppose if we have a word which is an open class word i.e. a noun or verb or adjective or adverb. Then it is a possibility that it might be assigned to multiple tags and we may face the ambiguity issue. HMM technique can overcome this issue to assign proper tags.[12].

1.7 Objective of Study

The main objective of this research is to present the mechanism for part-of-speech tagging in hindi language. Also to study all related work done till date in this field and

applying improvements to increase the accuracy of tagged words. The goal is to study all existing algorithms and implement selected algorithm for better accuracy in Hindi POS tagger. The aim is to built POS tagger in order to understand and analyse morphological structure of hindi language and improve the performance of tagger. Also to increase the tagset to improve the overall tagging accuracy.

The selected language is Hindi mainly because it is the most spoken language in India and not much work done in this field yet for this language.

1.8 Scope of Work

The scope of this report is upto literature survey of all related work and Implementation of basic POS tagger using Hindi WordNet and basic HMM technique. A brief summary of implementation done is added in later sections. Approaches which have been taken in building Hindi POS tagger has been discussed and so the existing algorithms in improving the accuracy. Proposed system is discussed in this report in later sections. Also method for assigning appropriate is discussed and implemented.

Chapter 2

Literature Survey

Hindi language is a free word order and very inflectional type of language. Whereas if we see language like English which is a fixed word order language, the relative positions of words depicts the relation between various parts of sentence. In Hindi, postpositions gives such relations, and inflecting nouns shows information about case, and similarly inflecting verbs reflects gender, number, and person information. Hindi can be represented by use of ASCII transliteration scheme for usage in commonly availble tools[4].

A large text is now there in electronic form in Hindi which is national language of India and hence, most spoken in India. To access this information, there is need of multi-lingual text information retrieval system is has increased. This has become active area of research nowadays. Earlier this area of work was mainly focussed on English Language. But in past decade, there is inclusion of Asian languages in this area. But Still, unavailability of tools and other lexical resources for languages like English and major European languages make it hard for the work in this area. This is very true for languages in Indian sub-continent[5].

2.1 Machine Aided Translation

2.1.1 Issues in parsing

Many problems exists which can cause problem is parsing if not handled well. Supreme intelligence is required to cope up with issues[2] like:

- Transformation of inanimate gender of English to Hindi Gender (any gender in Hindi)

- Next step After English Sentences or paragraphs are preprocessed, we get POS tagged English words. These words will be converted to Hindi words at word level. English-to-Hindi dictionary will be best choice for this. Hindi words we get after this, are organized in the S-V-O form to get the desired translation. Also after this, there is need to add **karaka** at right places to maintain structure of language

- If English word has various meanings then perform word sense disambiguation after POS tagging to deduce it to correct meaning according to the context. For example:
  Tarun mail the mail
  In above sentence, both 'mail' have different meaning. This problem is solved by POS tagging. First occurrence is verb and second is noun

- Some sentences give various meanings at sentence level. For Example:
  I hate annoying teachers
  In this cases, we need to select only one meaning

- Some sentences, after translating to Hindi, requires an subject to be implied. Example: **Catch the ball!** Here we can see that the Subject ('you') is added

2.1.2 Issue selected for research

The issue, POS tagging of Hindi corpus, is selected as main research issue. The aim is to build the POS tagger for Hindi corpus.

2.2 Stemming

Let us look at some stemmers build using different approaches in the past in table 2.1.

2.3 POS Tagger

2.3.1 Related Work

Many POS taggers are implemented in English language which use machine learning algorithms like decision trees, markov model, maximum entropy methods and many more. Some taggers are present which employ both stochastic and rule-based approaches. The average accuracy obtained using machine learning techniques is 93 to 98 %. English is

Table 2.1: Stemmers developed for different languages in the past

Author	**Technique Used**	**Model**	**Language**
Beth Lovins (1968)	Rule Based	-	English
Martin Porter (1980)	Rule Based	-	English
Wicentowski (2004)	Supervised Approach	Word Frame model	English
Brent(1995)	Information Theoretic Approach (Unsupervised)	Minimum Description Length (MDL) Framework	English
Brent and Snover (2001)	Unsupervised	Bayesian model	English and French
Goldsmith (2001)	Unsupervised	Minimum description length principle	English
Freitag (2005)	Unsupervised	Automatic clustering of words using co-occurrence information.	English
Bharati et al. (2001)	Observable paradigm Unsupervised	Morphological analysis and Generation	Hindi
Ganapathiraju (2006)	-	TelMore	Telugu
Ramanathan and Rao (2003)	-	Hand crafted suffix list and longest match stripping	Hindi
Majumder (2007)	clustering-based approach	YASS: Yet Another Suffix Stripper	Hindi
Pandey and Siddiqui (2008)	Unsupervised	Goldsmiths[7] Approach	Hindi
Patel et al. (2010)	Unsupervised	Goldsmiths[7] Approach	Gujarati

most used language in the world as well in day to day life and has very large corpora already , which gives freedom to use such machine learning methods. But not every language have such resources for example Hindi.[9].

Let see briefly all the related study undergone in morphology based disambiguation for Hindi POS tagger[9]:

- Bharati, in 1995, worked on a computational Paninian parser where tagging was done at parsing level
- Ray, in 2003, described an algorithm that groups Hindi words on basis of their tags. This results in reduction in the number of tags for a given sentence for some given

constraints on vocabulary categories that are possible for given sentence.

- There are many other online morphological taggers but their literature is non-existing.

A part-of-speech tagger and chunker for Hindi which uses maximum entropy framework is explained in [13] and tagger with hiddin markov model is discussed in [14]. Local word grouping in Hindi is given in [15]. POS Tagger developed till date are shown in table 2.2.

Table 2.2: POS taggers in the past

Author	**Approach**	**Year**	**Accuracy(%)**
Bharti et. al.	Morphological analyzer	1995	-
Singh et. al.	Decision tree based classifier	2006	93.45
Dalal et. al.	Pure maximum entropy based machine learning approach	2006	88.4
Shrivastava and Bhattacharya	Stemmer to generate suffixes	2008	93.12
Agarwal and Amni	Conditional Random Fields (CRF) with morphological analyzer	2006	82.67
Avinesh and Gali	Conditional Random Fields (CRF) with morphological analyzer	2006	78.66
Manju et. al.	HMM based tagger (Malyalam)	2009	-
Anthony et. at.	Support Vector Machines (SVM) (Malyalam)	2006	94.00

2.3.2 Challenges of Hindi POST

The inter-POS is ambiguity which is produced when a word has more than one possible tag. To understand this, lets us take an following example where *back* can be tagged in three ways[9]:

I get back to the back seat to give rest to my back

Such ambiguity will surely increase for a free order language like Hindi where a word has many variants. This is the why it is difficult for stochastic tagger in Hindi.

Intra-POS ambiguity is described as a word having one part of speech with different tag values.

One more difficult task is how to proper tag which defines both the form of word and the context used. Also new word always keep appearing in sentences or paragraphs. Therefore a new method is must for assigning the tag for such new words which are not

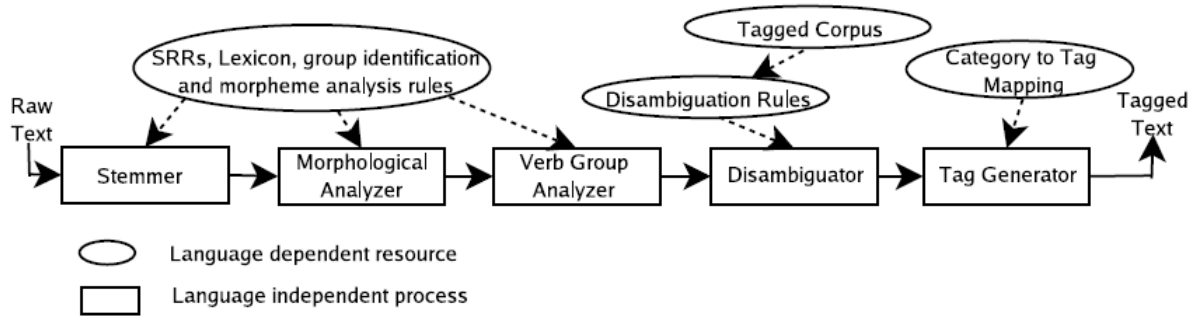

Figure 2.1: Overall Architecture of the Tagger

in vocabulary. This is achieved by using context information and the information by attaching some addition to base word, because these generated words in hindi language provides indication of a word's part of speech[9].

2.3.3 Design of Morphology Driven Tagger

Morphology driven tagger uses information of words which are created by addition to base words and tag it without using contextual information[9]. It also uses previous and next word in verb group to precisely detect the center verb and supplement words, then other POS categories are identified with the help vocabulary lookup of base word. The overall architecture of Hindi POS tagger[9] is displayed in Figure 2.1.

The lexicon which was used for comparing various POS tagging approaches[9], was taken from Hindi Wordnet as well as partial noun list from Anusaraka[9].

Now we will see approaches used to tag words in hindi language. The are Lexicon lookup based (LLB) approach, LLB with disambiguation(LLBD), Morphology-driven Tagger(MD), Baseline Tagging(BL) and Learning based (LB) tagger after 4-fold cross validation. However ambiguity remains in the tagged words and there is need of disambiguation rules. Below is the table showing average accuracy comparison[9] of various approaches in Figure 2.3.

The average overall accuracy brought up by augmenting morphological-driven tagger[9] with disambiguation rules is about 20 %. The rules can be generated from the training corpora by using machine learning algorithm like CN2 induction algorithm[16], HMM[12] etc.

Table 2.3: Average Accuracy(%) Comparison of various approaches

Approach	**Accuracy(%)**
Lexicon Lookup Based(LLB)	61.19
LLB with Disambiguation	86.77
Morphology-Driven(MD) Tagger	73.62
Baseline Tagging (BL)	82.63
Learning based(LB) with 4-cross Validation	93.45

The accuracy achieved using lexicon lookup based approach (LLB) is 61.19 % where as morphology-driven tagger was only bettered lookup approach by little margin but still contains lot of ambiguity. These results shows the need of using detailed morphological analysis. If a word is tagged in all possible forms then the accuracy of the tagger jumps to 73.62 %. Our aim is to keep best possible tag according to sentence's context and delete others disambiguation technique. The disambiguation task tags the word with most used tag for it. This technique is called as BaseLine (BL) tagging. The accuracy using this approach is 82.63% which better than approach taken by morphology driven tagger.

Limitation of BL is that no further improvement is possible there. On the other side, one can always find a way to improve MD tagger. It confirms that MD tagger works well for Verb Group and also with other close categories. There are around 30 % of the words which are either ambiguous or unknown. Thus, a disambiguation is must in this case. The graph[9] shows that the accuracy of POS tagged words is directly proportional to the size to corpus taken. It is also known as POS learning curve(Figure 4.3).

2.3.4 HMM based tagger

HMM based tagger assigns best tag to given input word by calculating the forward and backward probabilities of tags with the sequence[12]. This can be represented by following equation:

$$P(t_i \mid w_i) = P(t_i \mid t_{i-1}).P(t_{i+1} \mid t_i).P(w_i \mid t_i) \tag{2.1}$$

Here $P(t_i \mid t_{i-1})$ is the probability of a current tag given the previous tag and $P(t_{i+1} \mid t_i)$ is the probability of the future tag given the current tag. This captures transition between

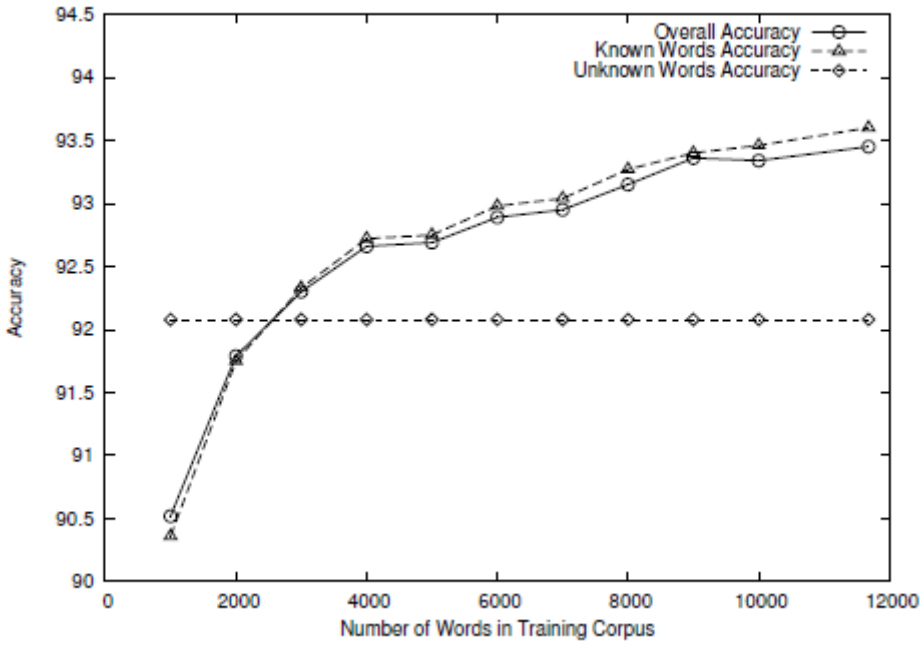

Figure 2.2: Accuracy versus Size of Corpus

tags. Also these probabilities can be calculated using:

$$P(t_i \mid t_{i-1}) = \frac{freq(t_{i-1}, t)}{freq(t_{i-1})} \tag{2.2}$$

Every tag transition probability is calculated as the ratio of frequency count of two tags coming together in corpus and the frequency count of the previous tag seen alone in corpus. This is usually done as it is well-known for some tags to precede other tags more. If we take an example, an adjective is most likely to be followed by a common noun but not by a postpositions or a pronoun.

Chapter 3

Proposed System

3.1 Existing System

While tagging the words in the Hindi language, there arise ambiguity where a word can be tagged in one or more categories. A human expert can easily determine these contexts and can predict a different POS tag to those words. Using HMM, this issue can be solved intuitively as the context of tags (before and after) are considered with respect to the current tag. This is the main advantage of using HMM which can assign the tag for a word as per the previous tag and the future tag of the word.

Using this statistical based HMM technique on training data, and along with disambiguating correct word-tag combinations using contextual information achieved 92.13% accuracy on test data.

3.2 Proposed System

The main aim is to built Hindi POS tagger. There will be Input document in Hindi language. These input token will then be preprocessed and refine by means of removal of stopwords and special characters. Then HMM technique will be applied on each word to predict the correct tag. Wordnet will provide required information about the Hindi word but will also help in identifying the words which impose ambiguity in tagging, i,e. word belonging to one or more category. This ambiguity will be resolved using HMM technique by using backward and forward probabilities and assigning the appropriate tag. The output will be Hindi tagged words (POS tagged). Also to try and implement tagging of new words added/overcome at runtime. Thus, Improving accuracy of the system. The

proposed system's flow chart is shown in figure 3.1.

As shown in flow chart, We will be reading file having sentences in Hindi language. Next step is tokenizing words. Before that we have to remove stopwords and special characters. After tokenizing Hindi words will be tagged, that is, Part of speech tagging will be done. This includes formation of rules for identification of Part of speech, number, punctuation and no tag[7]. Rules will be formed to mark the various types of tags to be done.

Further accuracy can be improved by increasing the tagset using Hindi grammar rules. Also to use techniques like "exploding input" or Longest suffix matching[17], Maximum Entropy Markov Model(MEMM)[18], etc to improve the tagging.

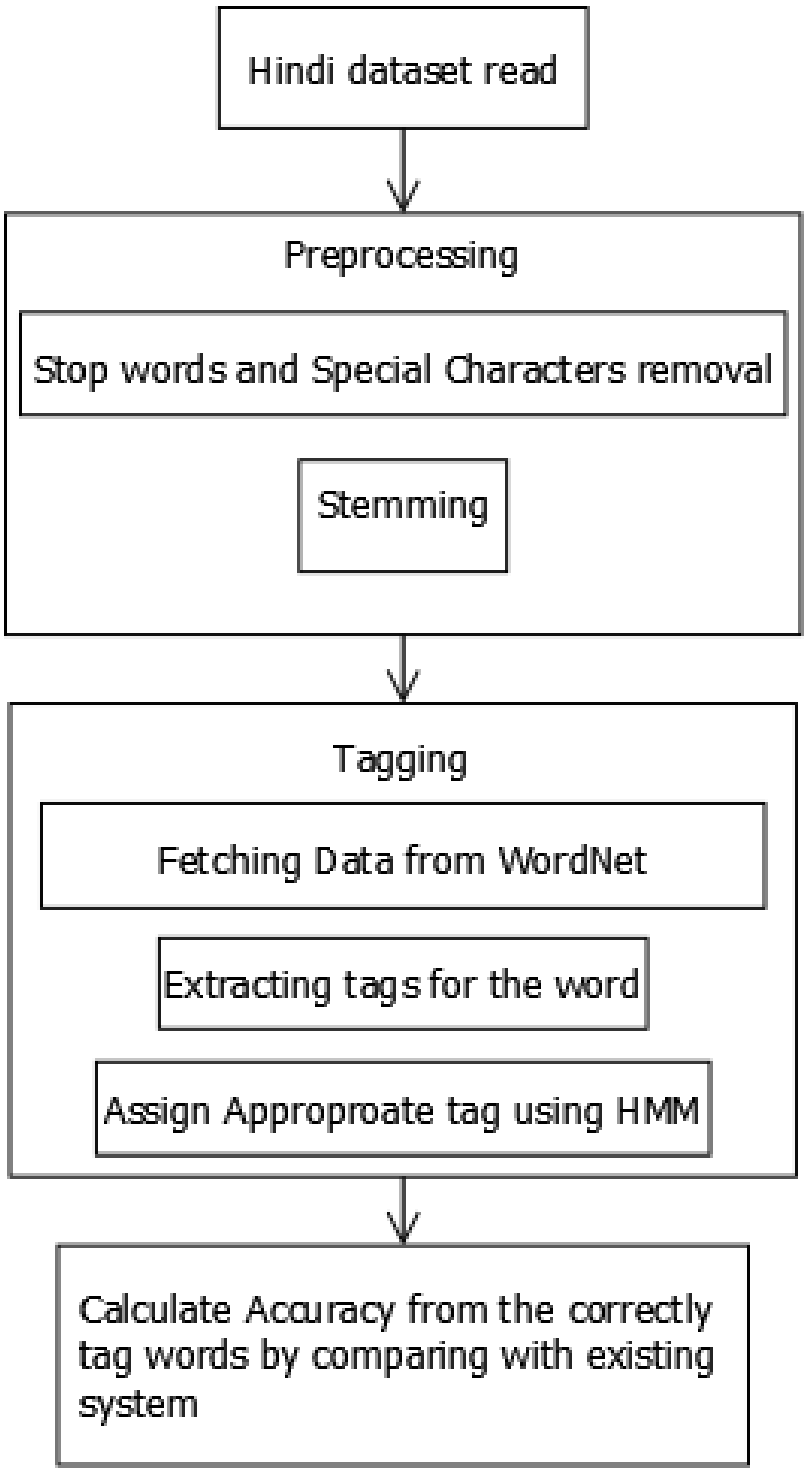

Figure 3.1: Flow Diagram of Proposed System

Chapter 4

Implmentation Details

POS tagger is implemented using Hindi WordNet. Main POS tags are identified using Hindi WordNet. Ambiguity is also identified. In this case, ambiguity in tags is then reduced using HMM statistical approach. The Tools used for implementation are:

- Eclipse
- Hindi WordNet
- Java Hindi WordNet Library(JHWNL)
- IIIT tagset
- Dataset from various domains (such as News, Tourism, Medical, etc.)

The current implementation includes part of speech tagging using Hindi Wordnet in Java and then tagging the appropriate tag using HMM model. Each word in input file will be analysed according to the WordNet database. If there is ambiguity in tagging, for example if word is having more than one tag, Forward and backward probabilities are calculated. The tag is assigned to the word according to these probabilities of tags along with the sequence of input. The input Hindi corpora taken is from Reddy[18]. The part of annotated Hinid corpora is shown in figure 4.1.

This input corpora is first preprocessed. In preprocessing, Hindi stopwords and special symbols are excluded. Then the output of preprocessed input is given to tagger input. Each word from the input is analysed using Hindi WordNet and respective tags are identified. The word with only one tag is marked and words with ambiguous tags are

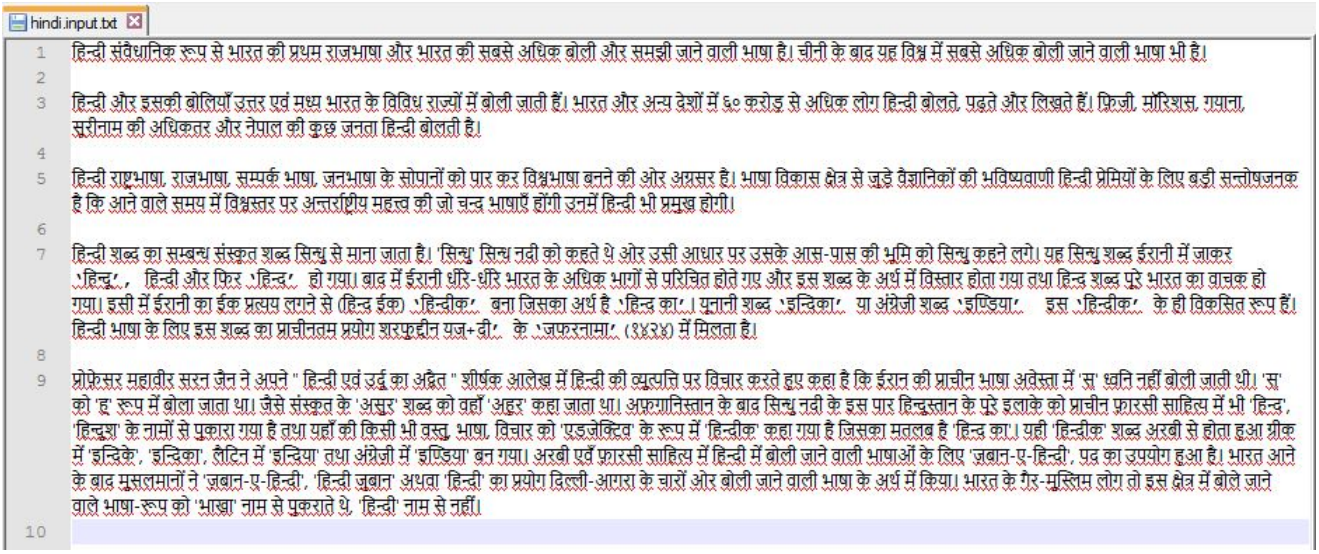

Figure 4.1: Part of Hindi Corpora given as input

identified. This words are tags using HMM technique. In this technique, backward and forward probabilities are calculated and then tag giving high probability is assigned.

Words with respective tags is stored in output file as well as printed in eclipse console. Each word is followed by its tag type. If unknown tag word is encountered, then the word is not tagged and next word is checked. The part of output is shown in figure 4.2.

16300 tokens were tagged from which only 12500 were correctly tagged. This means accuracy achieved is 76.0%. These tags are of only 4 types: Noun, Verb, Adjective, Adverb. To improve accuracy, tagset is increased. After increasing the tagset the accuracy increases. On comparing with popular approaches, The graph has been shown in Figure 4.3. This graph shows that our proposed tagger works better than Morphology-Driven and Lexicon Lookup based method but it still not good enough. Disambiguation aided POS tagger works better than this.

Figure 4.2: Part of output

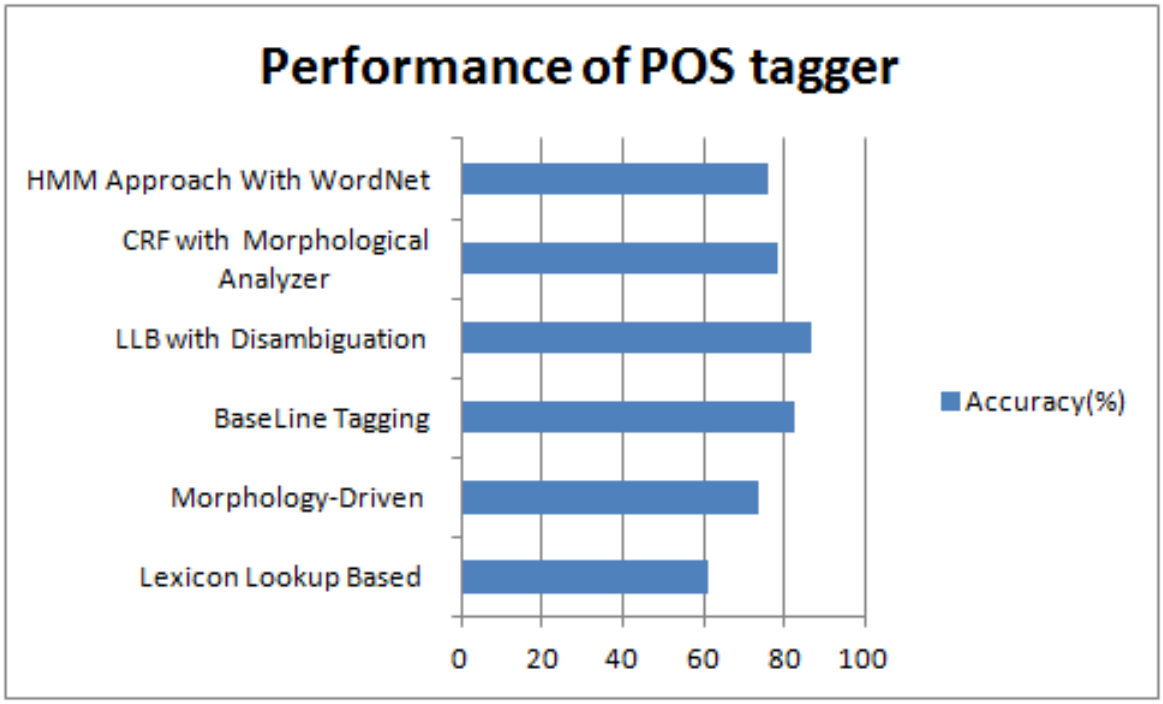

Figure 4.3: POS taggers Accuracy Comparison

Chapter 5

Conclusion and Future Scope

5.1 Conclusion

Hindi POS tagger has been implemented using Hindi WordNet and is able to tag input words. These Hindi sentences stored in file are taken as input and output is words with their POS tags. Those words which falls in more than one category are identified. Tags for words are derived with the help of Hindi WordNet and are assign correctly with HMM model. HMM model provide statistical method of calculation appropriate tag for the word by using probability mechanism. Further tagset is increased. Accuracy obtained from dataset of Tourism domain is 76%. It contained around 16300 tokens after cleaning and 12500 were tagged correctly. But Maximum accuracy which has been achieved is 83.26% using HMM approach[19] which is achieved after annotating the corpora for around 4 months. Also another large web corpora of Hindi is downloaded from Shiva, author of[20]. It was from health domain.

Also through literature survey following conclusion has been inferred. The maximum accuracy we get for Hindi POS tagging is 93.45% which uses Machine Learning Algorithm (LB Tagger). This tagger's accuracy has been increased using morphology based tagger. The Disambiguation rules are obtained by CN2 machine learning algorithm. The domain selected was news domain (BBC news). This approach concludes that augmenting Morphology-driven driver, the overall accuracy jumps upto 20%. A better preprocessing is need as well as intense use of contextual information is need to attain better accuracy and performance.

This proposed method can be further improved by adding disambiguation rules and Harnessing morphological richness of Hindi language. Better preprocessing with also improve the results. This approach gives good accuracy for words available in Wordnet but when new words are encountered, contextual information is needed with requires disambiguation rules.

5.2 Future Scope

Future work demands implementation of efficient Part of speech tagger by adding disambiguation rules. Better preprocessing with also improve the results. Stemming should be done to extract the morphological richness of Hindi language. Also increasing the tagset to classify into more discrete categories. Also using contextual information to increase accuracy and performance of tagger should be done. Also input can be preprocessed to achieve this. One way can be "explode input" in order to increase the length of input which will reduce the number of unique types of words encounter during learning[17].

References

[1] *Wikipedia, and other Internet resources.*

[2] T. A. R. K. B. V. M. W. Rekha S. Sugandhi, Ritika Shekhar, "Issues in parsing for machine aided translation from english to hindi," IEEE, 2011.

[3] L. B. M. K. K. K. P. K. Ronan Collobert, JasonWeston, "Natural language processing (almost) from scratch," Journal of Machine Learning Research 12 (2011), 2493-2537., 2011.

[4] D. D. R. Ananthakrishnan Ramanathan, "A lightweight stemmer for hindi,"

[5] T. J. S. Amaresh Kumar Pandey, "An unsupervised hindi stemmer with heuristic improvements,"

[6] P. B. Kartik Suba, Dipti Jiandani, "Hybrid inflectional stemmer and rule-based derivational stemmer for gujarati,"

[7] A. M. Nidhi Mishra, "Part of speech tagging for hindi corpus," International Conference on Communication Systems and Network Technologies, 2011.

[8] C. D. Manning and H. Schiitze, *Foundations of Statistical Natural Language Processing.* The MIT Press Cambridge, Massachusetts London, England.

[9] M. S. Smriti Singh, Kuhoo Gupta and P. Bhattacharyya, "Morphological richness offsets resource demand- experiences in constructing a pos tagger for hindi," 2006.

[10] A. Jain and D. KLobiyal, "A new method for updating word senses in hindi wordnet," IEEE, 2014.

[11] M. S. Mall and D. U. C. Jaiswal, "Developing a system for machine translation from hindi language to english language," 4th International Conference on Computer and Communication Technology (ICCCT), 2013.

[12] H. D. Nisheeth Joshi and I. Mathur, "Hmm based pos tagger for hindi," CS & IT-CSCP, pp. 341-349, 2013.

[13] U. S. S. S. Aniket Dalal, Kumar Nagaraj, "Hindi part-of-speech tagging and chunking : A maximum entropy approach,"

[14] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "Practical part-of-speech tagger," Xerox Palo Alto Research Center.

[15] S. S. A. B. Pradipta Ranjan Ray, Harish V., "Part of speech tagging and local word grouping techniques for natural language parsing in hindi,"

[16] T. N. Peter Clark, "The cn2 induction algorithm," Machine Learning Journal, 3 (4), pp261-283, Netherlands, 1989.

[17] M. Shrivastava and P. Bhattacharyya, "Hindi pos tagger using naive stemming : Harnessing morphological information without extensive linguistic knowledge,"

[18] U. S. Aniket Dalal, Kumar Nagaraj and S. Shelke, "Hindi part-of-speech tagging and chunking : A maximum entropy approach,"

[19] S. B. R and D. R. K. P, "Current state of the art pos tagging for indian languages - a study," IJCET,Vol. 1, Number 1, pp. 250-260, IAEME, May-June 2010.

[20] S. Reddy and S. Sharoff, "Cross language pos taggers (and other tools) for indian languages: An experiment with kannada using telugu resources," Proceedings of the 5th International Joint Conference on Natural Language Processing, pp. 11-19, November,2011.