Volume 2, Issue 1 January 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal

Of

Science, Engineering and Information Technology

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi, Urdu all research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: - Dr. Ashak Hussain Malik House No. 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

HINDI TO ENGLISH PART OF SPEECH TAGGER BY USING CRF METHOD

B.REVATHI[1] & DR.M.HUMERA KHANAM[2]

[1](M.Tech, Student, Dept. of Computer Science and Engineering, SVU College of Engineering, Tirupati, Andhra Pradesh, India.)

[2](Associate Professor, (M.Tech, PhD), Dept. Of Computer Science and Engineering, SVU College of Engineering, Tirupati, Andhra Pradesh, India.)

ABSTRACT

Part of Speech Tagging is the process of assigning a part of speech or other lexical class marker to each word in a corpus.POS tagging is a very important preprocessing task for language processing activities. This helps in doing deep parsing of text and in developing information extraction systems, semantic processing etc. The problem of tagging in natural language processing is to resolve the ambiguity, choosing the proper tag for the context. In this paper, we present a Condition Random Field (C.R.F) based Tagger for Hindi language. A sentence in Hindi language is given as input to the NLP tool the output is tagged sentence in which each word of the sentence is assigned with a unique part of speech tag. Accuracy is the prime factor in evaluating any POS tagger so the accuracy of proposed tagger is also discussed in this paper.

INDEX TERMS- POS, Tagging, CRF, Hybrid, Morphological analysis.

INTRODUCTION

Natural Language Processing (NLP) began in the 1950s as the intersection of artificial intelligence and linguistics. NLP is a fast growing technology at present and also it is a very important resource for fetching information from collections of huge amount of data with the help of imposing some queries and keywords. But there is problem of attractive information what the user exactly wants because it contains more than one document related to a particular thing, person or incident etc. For instance, when we search for some data in the warehouse with the help of some query, we may get a lot of un-important or irrelevant data instead of getting the exact data or information. So, in order to fetch the exact information from large collection of documents what the user exactly wants there is great need of some methods or mechanisms. This leads to the Information Extraction Research. Information Extraction (IE) is a method which helps in extracting the required or exact data. It is the process of fetching the required information from large collection of documents what the user want.

Morphology is the field of the linguistics that studies the internal structure of the words. Morphological analysis means taking a word as input and identifying their stems and affixes. Morphological Analysis is essential for Hindi it has a rich system of inflectional morphology as like other languages. Morphological Analyzer and generator is a tool for analyzing the given word and generator for generating word given the stem and its features (like affixes).

Part of speech tagging is the process of assigning a part of speech like noun, verb, preposition, pronoun, adverb, adjective or other lexical class marker to each word in a sentence POS tagging also known as Grammatical Tagging or Word Category Disambiguation. Hindi Language is considered morphologically rich and free order language so it is the biggest problem in Machine Translation, Language Learning & Teaching, Natural Language Generation, etc. to transfer correct part of speech to each word of a given input text depending on the situation. POS tagger plays important role to solve such problems. Many Hindi POS tagger at present available doesn't work properly and correct POS tagging in Hindi sentences because Morphophonemic changes are major problems in Hindi text. There are a number of approaches to implement part of speech tagger, i.e. Rule Based approach, Statistical approach and Hybrid approach.

A. Rule Based Approach

It uses linguistic grammar-based techniques to find tags. It needs rich and expressive rules and gives good results. It requires great knowledge of grammar and other language related rules.

B. Statistical Approach

1) Hidden Markov Model:

HMM is a generative model. The model assigns the joint probability to paired observation and label sequence and the parameters are trained to maximize the joint likelihood of training sets.

2) Maximum Entropy Markov Model:

It is a conditional probabilistic sequence model. It can represent multiple features of a word and can also handle long term dependency. It is based on the principle of maximum entropy which states that the least biased model which considers all known facts is the one which maximizes entropy.

3) Conditional Random Field Model:

It is a type of discriminative probabilistic model. It has all the advantages of MEMMs without the label bias problem. CRFs are an undirected graphical model (also known as random field) which is used to

calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

C. Hybrid Approach

Hybrid models are basically combination of rules based and statistical models. In Hybrid system, approach uses the combination of both rule-based and ML technique and makes new methods using strongest points from each method. It is making use of essential feature from ML approaches and uses the rules to make it more efficient.

APPLICATION OF POS TAGGING:

Partial parsing: Syntactic analysis

Machine Translation: POS tagger is playing important role in machine translation. This system is analysis of source language text to base on target language text.

Information Extraction: tagging a partial parsing help identify useful terms and relationships between them.

Information Retrieval: noun phrase recognition and query-document matching based on meaningful units rather than individual terms.

Question Answering: analyzing a query to understand what type of entity the user is understand what type of entity the user is looking for and how it is related to other noun phrases mentioned in the question.

SYSTEM DESCRIPTION

Many words are ambiguous in their part of speech. For example, "Book" can be a noun or a verb. However, when a word appears in the context of other words, the ambiguity is often reduced: in "Book that Flight" the word "Book" can only be a Verb. POS tagger is a system that uses context to assign parts of speech to words. Automatic text tagging is an important first step in discovering the linguistic structure of large text corpora. Part-of-speech information facilitates higher-level analysis, such as recognizing noun phrases and other patterns in text.

ALGORITHM FOR HINDI POS TAGGER:

We are design to following algorithm for Hindi POS tagger:

1) Read the input text and assign the same on a string type variable.

2) Breaking the text into sentences and further every sentence breaks into words and taking one by one word from the text and goes to next process.

3) Checking the word properties just like root/steam.

4) At this step we analysis the words root or steam existence. If it these exist then the go for tagging and applying derived rules otherwise go to the morphological analysis.

5) At this step we apply prefix, suffix on root/steam.

6) At this step we apply morphological synthesizer on word and tag the word.

7) If word has its correct GC then tag the word and display the result. Otherwise applying disambiguation rules for noun, verb, adjective etc. and tag the word.

8) Repeating the step from fourth to eighth for other words.

9) Output for all tagged words.

10) Stop

Examples

1) The input text is as follows:

जॉन एक अच्छा लड़का है

Then the output is Tagged sentence as follows

jon ek achchha ladaka hai

jon	- NN
ek	-NN
achchha	-JJ
ladaka	-NN
hai	-VB

2) The input text is as follows:

सीता पढ़ रही है

Then the output is Tagged sentence as follows

seeta padh rahee hai

seeta -NN
padh -VB
rahee -JJ
hai -VB

EXPERIMENTAL RESULTS

The POS tagging for hindi language using CRF takes the input as hindi text which is loaded into the dictionary first and then reads the input file by which the input text will be translated from hindi to english. Finally the translated text is divided into tokens and hence the output is the tagged text with full of POS details.

The snapshots for the whole procedure are as follows:

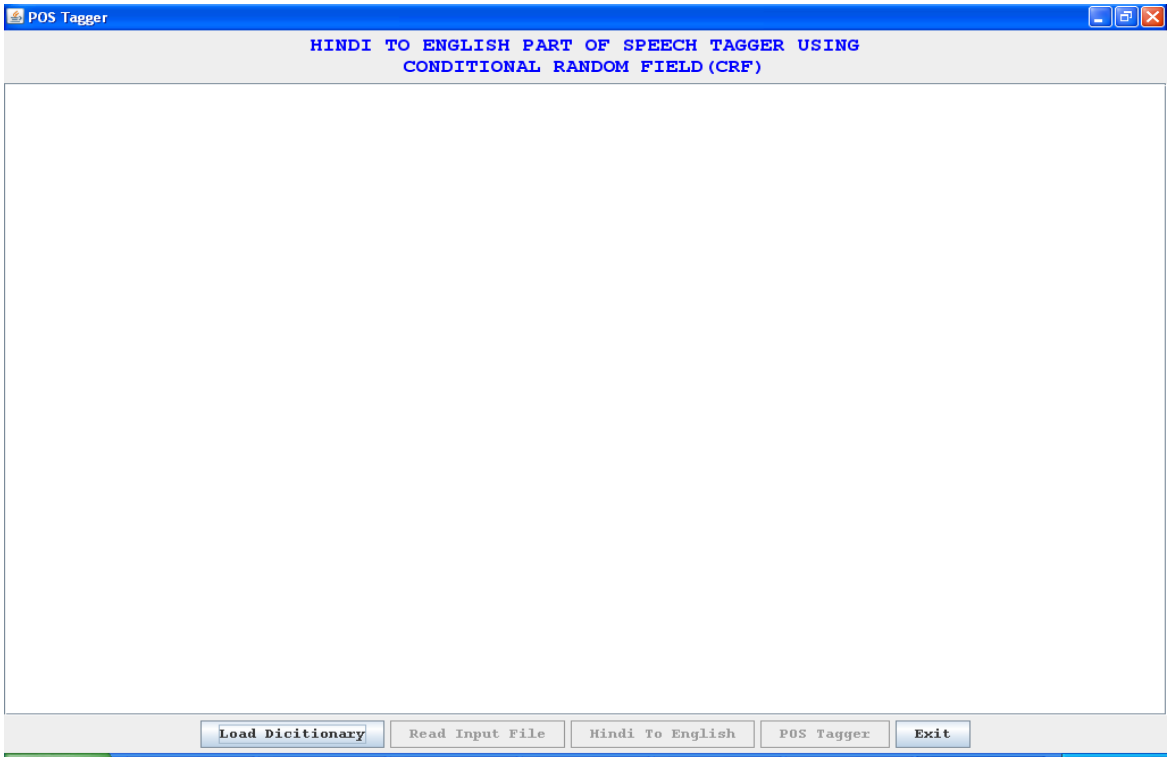

Fig: 1 Snapshot for "Load Dicitionary".

Fig: 2 Snapshot for "Read Input File"

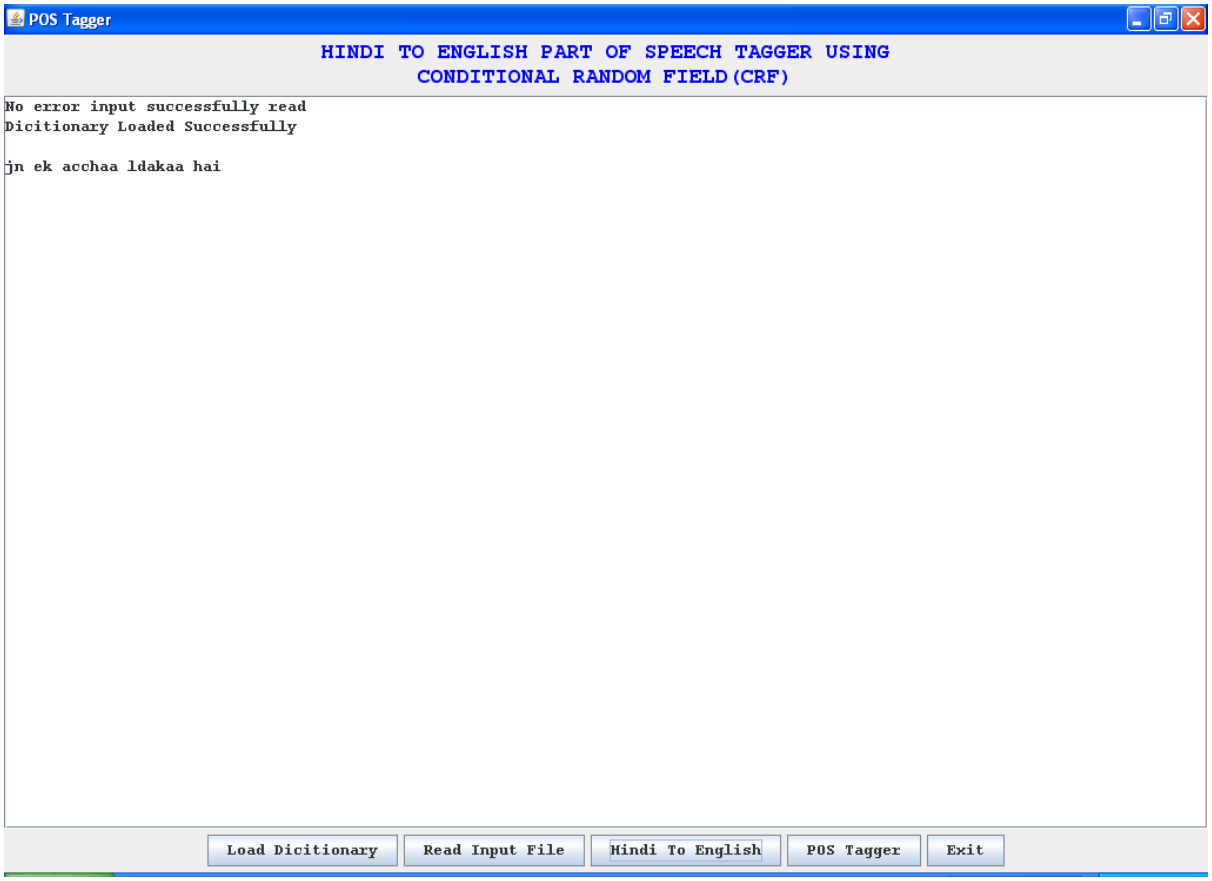

Fig: 3 Snapshot for conversion of Hindi to English

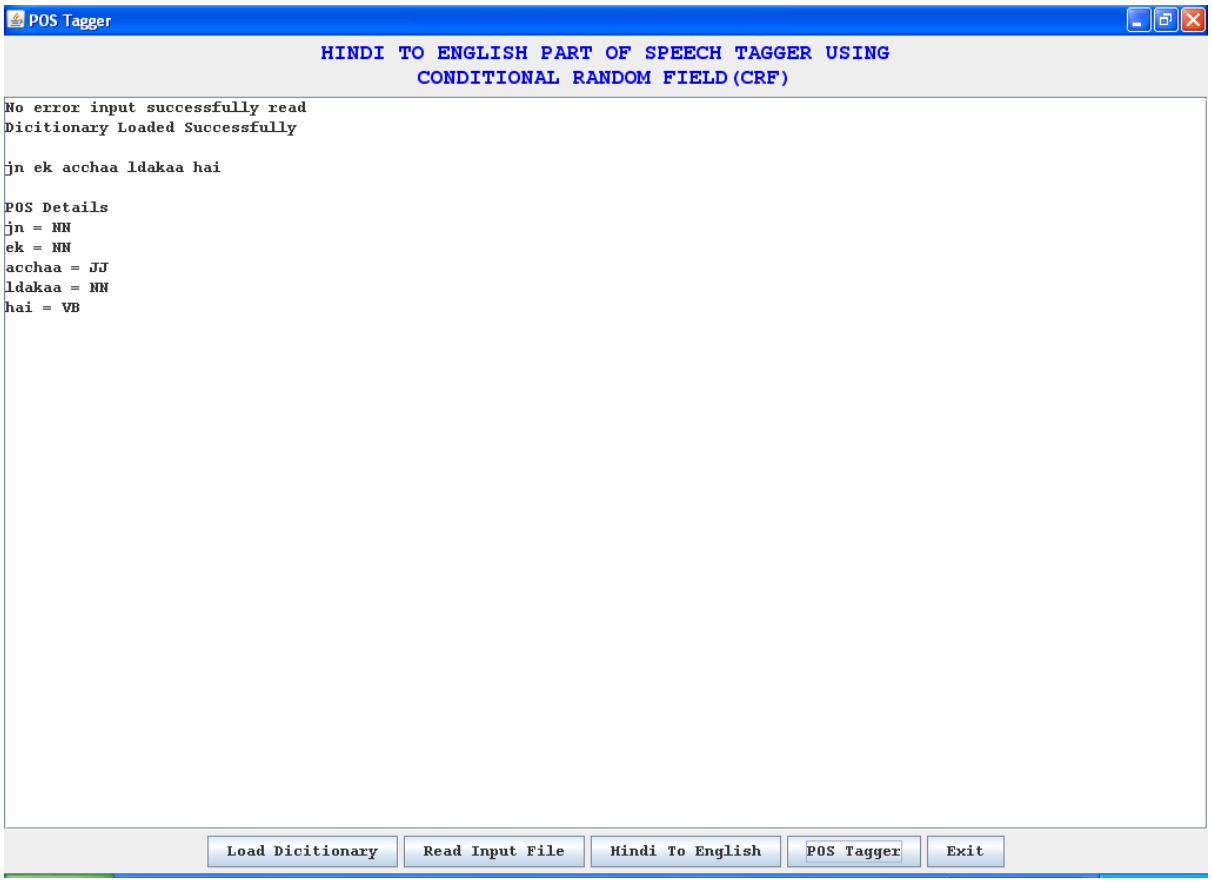

Fig: 4. Above output showing in English as 'john is a good boy'

Now by clicking on 'POS Tagger' button, it gives the POS details

CONCLUSION

In NLP POS tagging is the major task. When machine understands the text then it is ready to do any NLP applications. For that the machine should understand each and every word with its meaning and POS. Particularly in Machine Translation when the system understands the Pos of source language text, then only it will translate into target language without any errors. So POS plays such an important role in NLP.

In this POS tagging for Hindi language using CRF, the experiments were conducted with our corpus of 80,000 words. We have got an overall accuracy of 92%. In future developments of this work, it is intended increase accuracy for tagging models on the Hindi texts.

REFERENCES

- Uma Parameshwari Rao G, Parameshwari K: CALTS, University of Hyderabad, 'On the description of morphological data for morphological analyzers and generators: A case of Telugu, Tamil and Kannada'.
- Beesley, K. and L. Karttunen. 'Finite State Morphology'. Stanford, CA: CSLI Publications, 2003.
- Aduriz1, Agirre E., 'A word-grammar based morphological analyzer for agglutinative languages', University of lheBasqtloCotlnlry, Basque Country.
- Koskenniemi .K, Two –Level Morphology: A general Computational; Model for Word Recognition and Production, University of Helsinki, Helsinki,1983.
- Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", In Proceedings of the IEEE, 77 (2), p. 257-286February 1989.Available at: http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf
- Language: A Hybrid Approach" International Journal of Computational Linguistics (IJCL), Volume (2) : Issue (1) : 2011.Available at: http://cscjournals.org/csc/manuscript/Journals/IJCL/volume2/Issue1/IJCL-19.pdf
- Sanjeev Kumar Sharma and Gurpreet Singh Lehal. (2011). *Using Hidden Markov Model to Improve the Accuracy of Punjabi POS Tagger*, Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on June 2011, pp. 697-701.
- Dinesh Kumar and Gurpreet Singh Josan. (2010). *Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey*, International Journal of Computer Applications (0975 – 8887) Volume 6– No.5, September 2010, pp.1-9.
- Nidhi Mishra and Amit Mishra. (2011). *Part of Speech Tagging for Hindi Corpus*, In the proceedings of 2011 International Conference on Communication systems and Network Technologies, pp.554-558.
- Andrew Borthwick. 1999. "Maximum Entropy Approach to Named Entity Recognition" Ph.D. thesis, New York University.
- F. Jelinek. 1997. Statistical Methods for Speech Recognition. MIT Press.
- NavneetGarg, Vishal Goyal, SumanPreet. "Rules Based Part of Speech Tagger" in the proceedings of COLING 2012: Mumbai, December 2012.

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- Dr. Ashak Hussain Malik House No-221, Gangoo Pulwama - 192301 Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com