Parts Of Speech Tagging for Chhattisgarhi Language

[1]Manish kumar sinha, [2] Shubham kumar sahu, [3]Sachin ther
Department of Information Technology, Bhilai Institute of Technology, Durg, Chhattisgarh, India

Abstract :POS tagger is very much essential software that is used in creation of language translators and extraxtion of information .The problems in POS tagging in NLP(natural language processing) is finding how to tag each words in a given sentence. In this paper we are presenting Rule based POS tagger technique for Chhattisgarhi language so that this thing could further be used for developing parser and translators for the Chhattisgarhi language. Although Chhattisgarhi is new language and has various types (basically five) so the POS tagger can't be made common. Here we are mainly focusing on Chhattisgarhi language of central region of Chhattisgarh and with further improvements this could be used for others regions too but with separate methods and algorithms.

Keywords: POS, parts of speech, tagger, Chhattisgarhi, Natural Language processing

1. INTRODUCTION

POS (parts of speech) tagging comes under Natural Language processing in the field of computer science and information technology , which deals with the artificial intelligence i.e. how machine learns and defines ,process and shows the results based on the program coded in accordance to the algorithm created . POS is basically divided into eight parts and it's not very easy to tag all the times correctly with the help of algorithms created. The accuracy can be achieved more than 90% but achieving 100 % accuracy is next to impossible. Parts of Speech tagging is one the key building blocks (noun, pronoun, verb, demonstrative, etc) for developing Natural Language Processing applications. There are a number of techniques to implement POS tagger, i.e. Rule Based technique, Statistical technique and Hybrid technique In Rule-based tagger we design dictionary to assign the correct tags to each word in the sentences or file. Because of this tags dis-ambiguity can also be removed from the sentences. Statistical POS tagger is based on the frequency and probabilities of occurrences of particular tagged word. In which we use tagged data (or trained data) to give a tag to a untagged word. Hybrid based POS tagger is combination of Rule based technique and Statistical technique. In which we applied grammar rules tries to remove dis-ambiguity. POS Tagger helps in speech recognition, natural language parsing (NLP) and also in retrieval of information. Chhattisgarhi is new language so no related work had been done on this language earlier.

2. METHODOLOGY

To develop Chhattisgarhi POS tagger we use rule based technique and different standard POS tags are used that are given by Department of Ministry of Communications & Information Technology for Indian languages and some other tags like time, date and number tag. Adverb. Each main tag like noun , pronoun, verb , adverb and other various tag are further divided into various parts like noun is divide in two parts proper noun and common noun similarly all tags are divided into various parts. The working of system would be classified into two parts first finding the word in database and second if found than tag the word. The database for noun can be taken from Hindi only and others database like for pronoun , verb , adjective needs to be created with the help of persons having great knowledge of Chhattisgarhi. Now in else part of the second step if word is not present in the database then given below rules will be applied to it.

2.1 Algorithm

The algorithm for POS tagging as follows

1. Tokenize the words from a given sentence
2. Normalization of the output provided from first step like separating out punctuation marks and symbols
3. Searching of Numerical Tags like
   For Example: - 1996, 1-2, 1.2, 12 वव,२३, ६.७, ६-७ etc.
4. Searching of Dates
   For Example: - 08/06/1996, 21/04/1994, etc.
5. Searching of Abbreviations
   For Example: - ए.पी, आर.के . etc.
6. Search in Database and if found provide the tag accordingly
7. If not found in records then apply the given below rules and tag them
8. Show the tagged output to the user

2.2 Rules for identification of Tags

All the rules are probability based when discussing and getting knowledge from various language experts we have made this rules this rules are does not guarantee that this all will always be true but has a higher probability and possibility of being true

2.2.1 Find Noun rule

Rule 1: Any word next to adjective could be noun

For Example:-

ओ टूरा असली नचनिया हे!

In above example असली is adjective नचनिया is noun.

Rule 2: A word next to the relative pronoun could be noun

For Example:-

इही घर हरे जेला शुभम बनाये हे

In above example इही and जेला is relative pronoun and घर and शुभम is noun.

Rule 3: A word after the reflexive pronoun could be noun

For Example:-

ओ अपन डउका सन चल दिस

In above example अपन is reflexive pronoun and डउका is noun.

Rule 4: A word next to personal pronoun could be noun

For Example:-

ऐ मोर पुस्तक हे

In above example अपन is reflexive pronoun and डउका is noun.

Rule 5: A word previous to the post position could be noun

For Example:-

शुभम रायपुर में रथे

In above example रायपुर is noun and में is post position.

Rule 6: word before the verb could be noun.

For Example:-

मनीष नहाये बर गे हे

In above example मनीष is noun and नहाये is verb.

Rule 7: A word next or previous to noun could be noun.

For Example:-

शुभम रायपुर में रथे

In above example शुभम and रायपुर both are noun.

More rules can be produced for finding noun but this above are sufficient

2.2.2 Find Demonstrative rules

Rule 1: If word is pronoun in database and next word is also pronoun, then first word will be demonstrative.

For Example:-
ते कोन हरस.
In above example ते and कोन both are pronoun.

Rule 2: If current word is pronoun in database and next word is noun, then current word will be demonstrative.

For Example: -
ओ घर चलदिस
In above example ओ is pronoun and घर is noun.

2.2.3 Find Proper Noun Rules

Rule 1: If current word is not tagged and next word is tagged as proper noun, then there is high probability that current word will be proper noun.

For Example: - शुभम , मनीष
In above example शुभम and मनीष are proper noun

Rule 2: If current word is name and next word is surname then we tagged them as single proper name.

For Example: - अमन कुमार
In above example अमन is name and कुमार is surname.

2.2.4 Find Adjective Rules

Rule 1: there is very much chance that a word before a verb is adjective
सचिन तेंदुलकर बढ़िया खेलथे
In above example बढ़िया is an adjective and खेलथे is a verb

2.2.5 Find Verb Rules

Rule 1: If current word is not tagged and next word tagged as an auxiliary verb, then there is high probability that current word will be main verb.

For Example:-
ओ हा खाना खात रिहिस
In above example खात is main verb and रिहिस is auxiliary verb.

3. RESULTS AND DISCUSSION

Input Chhattisgarhi Sentence

शुभम कॉलेज में पढ़थे

Output

शुभम <N_NNP> कॉलेज< N_NN> में<PSP> पढ़थे< V_VM>

Several times this may show wrong output as it will show the results on the basis of database algorithms and rules that we have created but this may sometime vary from actual grammar or real POS tagging that a human may do

4. CONCLUSION

Our Research is Successful in tagging each and every word provided in the sentence . The accuracy is not 100% but is higher than 70% accuracy percentage varies that depends upon training of machine and training data as well as the testing data but it mostly above 70% , and if the sentence has all words that are available in the data base then the accuracy higher than 90% , this rule based tagging is not completely rule based as the tool used for finding from database used a statistical and condition random field approach , so the algorithm is rule based but the end product i.e. the software will be hybrid . In Chhattisgarhi language there are many issues like verb and post-position are combined in a single word most of the times , and most of the times its hard define gender by reading the sentences.

REFERENCES

1. Part-of-speech tagging , https://en.wikipedia.org/wiki/Part-of-speech_tagging
2. Rijuka Pathak and SomeshDewangan: Natural Language Chhattisgarhi: A Literature Survey, *International Journal ofEngineering Trends and Technology (IJETT) – Volume 12 Number 2 - Jun 2014*
3. Technology Development for Indian Languages (TDIL), http://tdil-dc.in
4. Natural language processing https://en.wikipedia.org/wiki/Natural_language_processing
5. Navneet Garg, Rule Based Hindi POS Tagger, Proceedings of COLING 2012: Demonstration Papers, pages 163–174,COLING 2012, Mumbai, December 2012.
6. Eric Brill. (1992). A Simple Rule Based POS Tagger, In Proceeding of the Third Conference on Applied Computational Linguistics (ACL), Trento, Italy, 1992, pp.112–116.
7. Dinesh Kumar and Gurpreet Singh Josan. (2010). POS Taggers for Morphologically Rich Indian Languages: A Survey, International Journal of Computer Applications (0975 – 8887) Volume 6– No.5, September 2010, pp.1-9.
8. Antony P J and Dr.Soman K P. (2011). POS Tagging for Indian Languages: A Literature Survey, International Journal of Computer Applications (0975 – 8887) Volume 34– No.8, pp.22-2
9. Vikas Pandey and Shubham Kumar Sahu in BITCON (feb2017) :Survey paper of various POS tagging techniques for Hindi Language.
10. Neetu Aggarwal and Amandeep kaurRandhawa(BGET Sangrur): Survey on POS tagging for Indian languages,International Journal of Computer Applications (0975 – 8887).
11. Jyoti Singh, Nisheeth Joshi and ItiMathur (Department of Computer Science Banasthali University Rajasthan, India): Devlopment of Marathi POS tagger using Statistical approach .
12. Er.Davinder Kaur and Er.UbeekaJain : Automatic Rule detection and POS tagging in Punjabi text, International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 6 Issue 3 March 2017, Page No. 20780-20784.
13. Pravesh Kumar Dwivedi and Pritendra Kumar Malakar: Hybrid Approach Based POS Tagger for Hindi Language, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) August 2015, PP 63-68 ,ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online).