International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-4, September 2018

Rule Based Parts of Speech Tagger for Chhattisgarhi Language

Vikas Pandey, M.V Padmavati, Ramesh Kumar

Abstract*: *There is an increasing demand for machine translation systems for various regional languages of India. Chhattisgarhi being the language of the young Chhattisgarh state requires automatic languages translating system. Various types of natural language processing (NLP) tools are required for developing Chhattisgarhi to Hindi machine translation (MT) system. In this paper, we are presenting rule based parts of speech tagger for Chhattisgarhi language. Parts of Speech tagging is a procedure in which each word of sentence is assigned a tag from tag set. The Parts of Speech tagger is based on rule base which is formed by taken into consideration the grammatical structure of Chhattisgarhi language. The system is constructed over corpus size of 40,000 words with tag set consists of 30 different parts of speech tags. The corpus is taken from various Chhattisgarhi stories. The system achieves an accuracy of 78%.

Index Terms: Chhattisgarhi, Machine Translation, Natural Language Processing, Parts of Speech tagger, Rule Based System.

I. INTRODUCTION

Most of the regional languages are low resources language. Some Indian languages are called low resource language as grammatical rules and literary work related to these languages is not present in public domain. Pre-processing task like POS tagging is a challenging task for these languages. In POS tagging process a specific grammar class which is called as tag is assigned to a word in the sentence from tag set. Tag set is a collection of grammar class which consist of English abbreviations like N(Noun),VM(Verb), PP(Preposition) etc.[1]. Parts of Speech (POS) tagging is a process of identifying the suitable class of tag for a word from a given tag set. It is very important task of pre-processing activity in machine translation. Machine translation systems take a source language and convert it into target language. Various tools are required in machine translation systems like tokenize, POS tagger, morphological analyzer and parser. POS tagger comes under pre-processing phase of machine translation system. Most of the regional languages are low resources language. Some Indian languages are called low resource language as grammatical rules and literary work related to these languages is not present in public domain. Pre-processing task like POS tagging is a challenging task for these languages. In POS tagging process a specific grammar class which is called as tag to a word in the sentence from tag set. Tag set is a collection of grammar class which consist of English abbreviations like NN (Noun), VM (Verb), PP (Preposition) etc.[2].

Example 1: हमन □□□□ □□□□□□□ म □□□□□□ □□□□□

WORDS	हमन	दुनो	बैलगाड़ी	म	रायपुर	जाबों
TAGS	PRP	N	N	PP	N	VM

Table (a): Chhattisgarhi words and its tags taken from Chhattisgarhi tag set.

There are various approaches for POS tagging: Rule based approach, Statistical approach and Hybrid approach [2, 3]. Accuracy factor is the most important factor in deciding the performance of POS tagger [2].

The Rule Based POS tagging approach is based on grammar rules that are framed by observing the grammatical structure of any language. These rules can be written in form of production grammar rules. Example:

"A proper noun is always followed by a noun" as in the Table (a) हमन (Pronoun) is followed by □□□□ (Noun)

There are some limitations of rule based approach; the main limitation is the formation of rule base. In this a rule is formulated for each condition [2, 3].

The Statistical Based POS tagging approach is based on two important factors. These are: Frequency and probability of occurrence of any word .In this approach most frequently used tag for a specific word in the annotated training dataset is used to tag that word in the un annotated dataset .The limitation of this system is that some sequences of tags can come up for sentences that are not correct according to the grammar rules of a certain language [3].

In Hybrid Approach the probability theory of statistical method is used to train the corpus and then the set of production rules are applied on the testing corpus for tagging of testing corpus [2, 3]. POS tagging process is broadly classified into two models: Supervised Model and Unsupervised Model [4]. Classification of POS Tagging is shown in Figure 1.

Revised Manuscript Received on 30 September 2018.
***** Correspondence Author
Vikas Pandey, Dept. of Information Technology, Bhilai Institute of Technology, Durg, India
Dr. M.V Padmavati, Dept. of Computer Science and Engg., Bhilai Institute of Technology, Durg, India
Dr. Ramesh Kumar, Dept. of Computer Science and Engg., Bhilai Institute of Technology, Durg, India

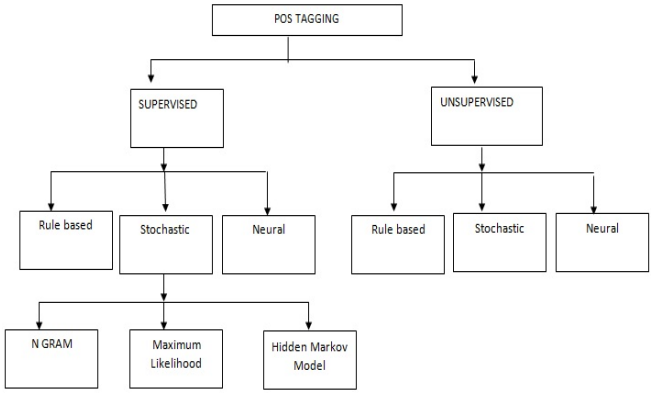

Figure 1: Classification of POS Tagging

II. LITERATURE SURVEY

Research has already been done in morphologically rich Indian languages like Hindi, Bengali, Telugu, Marathi, Tamil, Urdu, Gujarati, Kannada, Malayalam, Odia and Punjabi. There are some low resource languages in India like Awadhi, Magahi, Nimadi, Bhojpuri, and Chhattisgarhi for which machine translation tools have not been developed yet.

A POS tagger was developed using conditional random field for Bengali language In this system contextual information of the words has been used to search different POS tags for various tokenized words. The system was evaluated over a corpus of 72,341 words with 26 different POS tags and system achieved the accuracy of 90.3% [5].

A POS tagger was developed using Hidden Markov Model for Hindi, which uses a Naïve stemmer as a pre-processor based on longest suffix matching algorithm to achieve accuracy of 93.12% [6].

A POS tagger was developed using Hidden Markov Model for Assamese. Unknown words were tagged using simple morphological analysis .The system was evaluated over a corpus of 10,000 words with 172 different POS tags and system achieved the accuracy of 87% [7].

A POS tagger was developed using Hidden Markov Model for Hindi .They uses Indian language POS tag set to develop this tagger and achieved the accuracy of 92% [8].

III. METHODOLOGY

This system is developed using rule based approach and tagging will be done by the help of POS tag made in consultation with Chhattisgarhi linguistics expert. The system mainly works in two steps-firstly the sentences are spitted into words by the help of line splitter program and input words are found in the database; if it is present then rules are applied to tag to assign if a proper tag and if it is not found then UNK tag will be assigned to it.

3.1 Algorithm

The algorithm used for rule based part of speech tagger for Chhattisgarhi language is as follows:

1. Chhattisgarhi sentence is tokenized by the help of line splitter program.
2. The Tokenized Chhattisgarhi words are normalized.
3. Search the numbers and tag them by using regular expression.
4. Abbreviations are searched using regular expression.
5. In the database all the input words are searched and tag the word according to appropriate tag.
6. UNK tag will be given to the unknown words.
7. The tagged words are then displayed.

3.2 Rules that are Applied to Identify Different Tags

3.2.1 Noun Identification Rules

Rule 1: If word is adjective then there is high probability that next word will be noun.

For Example:-

□□□□ □□□ □□ □□□□

In above example □□□□ is adjective □□□ is noun.

Rule 2: If word is relative pronoun then there is high probability that next word will be noun.

For Example:-

□□□ घर □□□ □□ □□ □□□□ □□□□□ □□ ।

In above example □□□ and □□□□ is relative pronoun and घर and □□□□ is noun.

Rule 3: If word is reflexive pronoun then there is high probability that next word will be noun.

For Example:-

ओअपन □□□ सन चल □□□□

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

Rule 4: If word is personal pronoun then there is high probability that next word will be noun.

For Example:-

ऐ □□□ □□□□□□ □□□

In above example □□□ is personal pronoun and □□□□□□ is noun

Rule 5: If current word is post position then there is high probability that previous word will be noun.

For Example:-

□□□□□□ □□□□□□ म □□□□

In above example □□□□□□is noun and □□□ is post position.

Rule 6: If current word is verb then there is probability that previous word will be noun.

For Example:-

□□□□ □□□□□ बर □□ □□ ।

In above example □□□□ is noun and □□□□□ is verb.

Rule 7: If word is noun then there is probability that next or previous word will be noun.

For Example:-

सुरिवात बिलासपुर म पढ़ते ।

In above example □□□□□□□ and □□□□□□□□ both are noun.

3.2.2 Demonstrative Identification Rules

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

For Example:-

□□ □□□ हरस ।

In above example current word is □□□ and next word is हरस and both are pronoun so □□ is demonstrative.

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

For Example: -

ओघर चल □□□ ।

In above example current word is घर which is noun and next word is चल which is a verb, so ओis demonstrative.

3.2.3 Proper Noun Identification 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: -

□□रण , □□□□□

In above example □□□□ and □□□□□ are proper noun.

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

For Example: -

□□□□□□□ □□□ □□□□ will be tagged as □□□□□□□□□□ □□□□ where '□□□□□□□□□□' is proper noun.

3.2.4 Adjective Identification Rules

Rule 1: There is more chance that a word before a verb is adjective.

□□□□ □□□□□□□□ □□□□□ □□□□□□

In above example □□□□□ is a adjective and □□□□□ is a verb

3*.2.5 Verb Identification 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□□ □□□□ □□□ □□□□□ ।

In above example □□□ is main verb and□□□□□is auxiliary verb.

IV. RESULTS AND DISCUSSION

In order to test the system, few lines of a Chhattisgarhi story titled '□□□□□□' is taken a input sentence :

Input Chhattisgarhi Sentence

आज□□□□□□□□ □□ न □□□□ □□□□ □□□□□ □□ □□□□□ ह □□□□□ भर □□ □□□□ हवय।

Output Chhattisgarhi Sentence

Chhattisgarhi Words	Tagging
आज सुरुत्ती हे न तउने पाके दियना के अंजोर ह रइपुर भर मं बगरे हवय।	आज <N_NN > सुरुत्ती <V_VM > हे <V_VAUX>न <PSP>तउने <R_PRP>पाके<CC_CCD> दियना<N_NN>के<PSP>अं जोर<N_NN>ह<CC_CCD> रइपुर<V_VM> भर<V_VAUX>मं<V_VAU X>बगरे<CC_CCD>हवय<V
	_VM>।<RD_PUNC>

Table1: Chhattisgarhi words and corresponding tags

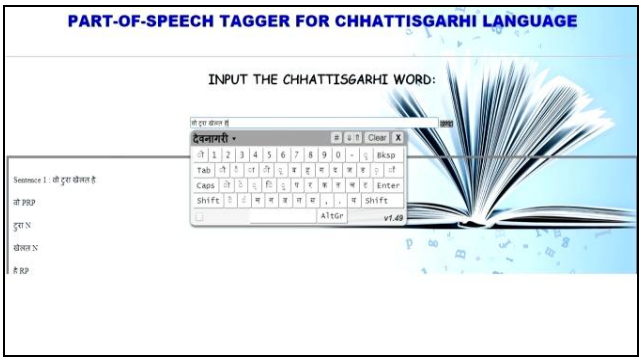

Figure 2: POS Tagging of Chhattisgarhi sentence

V.CONCLUSION AND FUTURE WORK

In this paper Part of Speech tagger for Chhattisgarhi language by the help of using rule based technique has been discussed. The sentence is broken in to tokens by line splitter program The tokenized words are search in the database and then appropriate rules are applied to tag them. The system is constructed over corpus size of 40,000 words with tag set consists of 30 different parts of speech tags. The test data is taken from various Chhattisgarhi stories and system achieves an accuracy of 78%.

The main limitation of rule based part of speech tagger is that it completely depends on rule. If rule is not present for a word then it will not be tagger by the system due to this accuracy of system will decrease.

In future there is need to shift towards neural network based system so that test data can be automatically gets trained and by increasing the size of corpus the accuracy of system will also increase.

REFERENCES

1. Agrawal, R., Ambati, B., & Singh, A.Singh.(2012). A GUI to Detect and Correct Errors in Hindi Dependency Treebank. In Proc.of Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, 1907-1911.
2. Hasan, M., F.,Uzzaman, N., & Khan, M.(2006). Comparison of Different POS Tagging Techniques (n- grams, HMM and Brill's Tagger) for Bangla. International Conference on Systems, Computing Scienc and Software Engineering (SCS2 06) of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering.
3. Kumawat, D., & Jain,V. (2015). POS Tagging Approaches: A Comparison. International Journal of Computer Applications ,vol 118,32-38
4. Antony, P.,J.(2011). Parts Of Speech Tagging for Indian Languages: A Literature Survey. International Journal of Computer Applications, Vol 34, 22-29.
5. Ekbal, A., Haque, R, & Bandhopadha, S. (2007). Bengali part of speech tagging using Conditional Random Field. In Proc. of SPSAL2007. 131-136.
6. Shrivastav, M., & Bhattacharyya, P. (2008). Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge. In Proc.of ICON, 1-6
7. Sharia, N., Das, D., Sharma U. , Kalita, J. (2009). Part of Speech Tagger for Assamese Text . In Proc. of the ACL-IJCNLP, 33-36.
8. Joshi, N., Darbari, H., & Mathur, I. (2013). HMM Based POS Tagger For Hindi. In Proc.of CCSIT, SIPP, AISC, PDCTA, 341-349