Developing Hindi POS Tagger for Homoeopathy Clinical Language

Pramod P. Sukhadeve and Sanjay K. Dwivedi

Department of Computer Science
Babasaheb Bhimrao Ambedkar University, Lucknow, India
sukhadeve.pramod@gmail.com, skd200@yahoo.com

Abstract. Part of speech tagging is one of the most basic preprocessing tasks of machine translation in NLP. The problem of tagging in natural language processing is to find a way to tag every word in a text as a meticulous part of speech. In this paper, we first present different approaches and some of the grammatical rules for tagging homoeopathy clinical sentences. Further in the paper we have our approach development of a Hindi tagger by using homoeopathy clinical sentences, for this purpose we have developed a corpus comprising of 250 sentences at present having 20060 words and 3420 tokens. The accuracy of POS tagging is calculated by using standard formula, and achieved the accuracy of 89.55%.

Keywords: POS tagging, Grammar rules, Homoeopathic Corpus, clinical words, POS Approaches.

1 Introduction

Homoeopathy comprises the harmless organism of medicine for treating maladies such as Keloids, Anaemia, Migraine, Rheumatism headache, blood pressure and also for urgent situations. Homoeopathy thus signify an awfully successful way of treatment. Doctors usually write prescriptions and reports in English which is unable to understand by most of the peoples because in India Hindi is widely spoken language. This is the major problem of communication between doctor and patient. So, to remove communication gap between doctor and patient we are endeavor to develop a machine translation system in which part of speech tagging is one of the step.

Part-of-speech tagging is the one of the most indispensable problem of NLP. It is a technique for assigning correct part of speech to each word of a given input sentence depending on the context. The significance of part-of-speech for language processing is the hefty amount of information they give about a word. POS tagging can be used in Text to Speech, Text to Text, Information retrieval, shallow parsing, Information extraction, Linguistic research for corpora [1] and also as an intermediate step for higher level NLP task such as parsing, semantics, translation and many more [2]. Thus POS tagging is a necessary application for advanced NLP application in Hindi or any other languages.

N. Meghanathan et al. (Eds.): CCSIT 2012, Part III, LNICST 86, pp. 310–316, 2012.

We start this paper by giving an overview of a few POS tagging, different approaches of POS tagging, grammatical rules of POS tagging, and then we describe the proposed POS tagging model with example and results obtained.

2 Approaches to POS Tagging

There have been many implementations of POS tagger using several techniques of machine translation, mainly for Corpus-rich languages like English. Such as, transformation-based error-driven learning based tagger [3] in this paper, E. Brill describe a rule based approach to automated learning of linguistic knowledge, maximum entropy Markov model based tagger [4] this is a new Markovian sequence model which is closely related to HMMs. A POS tagger for English based on probabilistic triclass model was developed [5]. A statistical POS tagger TnT proposed by [6] based on Markov models with a smoothing technique and methods to handle unknown words. POS tagging is typically accomplished by rule-based systems, probabilistic data-driven systems, neural network systems or hybrid systems. For languages like English or French, hybrid taggers [7] have been able to achieve success percentages above 98%. Another tagger for Malayalam was proposed [8] which is based on machine learning approach with Support Vector Machine (SVM) [9]. The objective was to develop a tag set appropriate for Malayalam.

Another approach for POS tagging is based on incorporating a set of linguistic rules in the tagger. A comparison between stochastic tagger and tagger [10] build with handcrafted linguistic rules which explains about the ambiguity, the error rate of the statistical tagger is greater than that of the rule-based. Some implementations combine the statistical approach with the rule-based, to build a hybrid POS tagger. Such a tagger was constructed by [11] for Hungarian, which shares many difficulties such as free word order, with Hindi language.

Due to non-availability of statistical information in Hindi, purely rule-based systems are only able to solve the problem of POS tagging. Such systems will eliminate the large number of definitely wrong tagging which would otherwise be present if no constraints were present. The partial POS tagger for Hindi presented the reduced error rate of possible tagging for a given sentence by imposing some constraints on the sequence of lexical categories that can typically occur in a Hindi sentence.

3 POS Tagging Rules

For the task of Hindi POS tagging, some of the rules are listed below:

I) Noun:

Nouns in Hindi are inflected for gender, number, and case. There are three declensions of nouns;

Declension 1 includes आ [*Aa*] at the end of masculine nouns.

Declension 2 includes all other masculine nouns, and

Declension 3 includes all feminine nouns.

There are two genders in Hindi: masculine and feminine. The gender of a large number of unresponsive nouns can be predicted by their endings, there are no fixed rules for assigning the genders. We can make some general observations as follows.

i) Most of the आ [*Aa*] ending masculine nouns have their feminine forms ending in ई [*Ee*].

Masculine	Feminine
लड़का [*Larkaa*] (boy)	लड़की [*Larkii*] (girl)
बच्चा [*Baccha*] (child)	बच्ची [*Bacchi*] (child)

in above example suffix of masculine is आ [*Aa*]and suffix of feminine is ई [*Ee*]. The final – आ [*Aa*] in the masculine nouns is replaced by – ई [*Ee*] in their feminine forms.

ii) Most of the –ई [*Ee*] ending animate masculine nouns have their feminine forms ending in –अन [*Aan*].

Masculine	Feminine
धोबी [*dhobee*] (laundress)	धोबन [*dhoban*] (laundress)

iii) Some nouns ending in –आ [*Aa*] form their feminine by replacing –इया [*Eya*].

Masculine	Feminine
डिब्बा [*diibaa*] (Box)	डिब्बियाँ [*dibeeyaan*] (Boxes)

iv) The suffix –नी [*Nee*] is added to the masculine nouns to form the feminine.

Masculine	Feminine
डॉक्टर [*daaktar*] (Doctor)	डॉक्टरनी [*daaktarnee*] (Doctor)
मास्टर [*maaster*] (Sir)	मास्टरनी [*maasternee*] (Madam)

II) Main Verbs

There are three types of main verbs: simple verbs, conjunct verbs, and compound verbs. A simple verb may consist of one main verb and person, gender, number, tense and aspect markers. In the compound verb construction, the person, gender, number and aspect markers are taken by the explicators/operators, and in the conjunct verbal construction they are taken by the verb element. We will classify the verbal constructions as intransitive, transitive, intransitive, causative, dative, conjunct and compound.

i) Intransitive Verbs

Intransitive verbs like आ [*Aa*], जा [*Jaa*], बैठ [*Baith*], do not take a direct object and are not marked by any postposition in the present or future tense. Subjects in such cases are controlled by the verb agreement.

वह जाता है । [*wah jaata haai*] (he goes)
शाम अस्पताल जाएगा। [*shaam aspataal jaayegaa*] (Shyam will go to hospital)

Besides verb agreement, subjects demonstrate a number of other properties which are explained below. Intransitive verbs in the past tense take their subjects in the direct case.

वह बहुत थक गई। [*wah bahoot thak gayee*] (she is very tired).
डॉक्टर समय पर आया। [*daaktar samay per aaya*] (Doctor came on time).

Some intransitive verbs, such as खेल [*khel*] (game/play) and पढ़ [*pad*] (study) may sometimes be used as transitives when they take abstract nouns as objects.

डॉक्टर ने किताब पढ़ी। [*daaktar ne kitaab padii*] (Doctor read a book)
डॉक्टर बोला। [*daaktar*] (doctor said).

ii) Transitive Verbs
Transitive verbs, such as इलाज [*Eelaaj*], दे [*dye*], कर [*kar*], take direct objects and in the past tense they require their subjects with the object in gender and number.
डॉक्टर ने मरीज का इलाज किया । [*daaktar nye mariij ka illaj kiya*] (doctor treated the patient)

4 POS Tagging Model

The proposed tagger for Hindi language has 29 tags where there are 5 tags for Nouns, 1 tag for Pronoun, 7 tags for Verbs, 3 tags for Punctuations, 1 for each adjective, adverb, conjunction, reduplication, intensifier, postposition, emphasize, determiners, complimentizer and question word. The proposed architecture for POS tagging is based on the model of transformation based tagger :

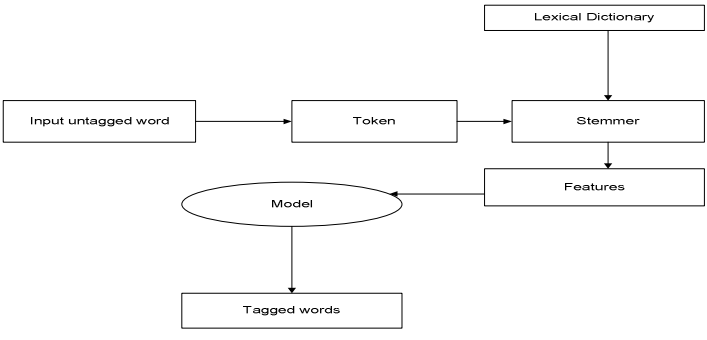

Fig. 1. Proposed model of POS tagger

The POS tagging model consist of different modules which accomplish different functionalities for better accuracy of POS tagger. In this model we first input the untagged text, which is further tokenize and then stemmed by lookup into the lexical dictionary and then with the help of tagging rules, each word is conceded to model where course of action is taken for tagging and in the output we acquire Hindi tagged words.

5 Tagging Example

Part of speech tagging problem is defined as the task of providing the correct grammatical information for words in sentences. We try manually some of the Hindi Language sentences for tagging which are as follows:

Input: राज दवा खा रहा है । [*Raaj dawa kha raha hai*] (Raj is eating medicine*)*
Output: राज_NP दवा_NP खा_VB रहा_DM है_DM

Whereas, NP- is Proper Noun
VB- is Verb
DM-is Determiner

6 Result Analysis

The precision of any part of speech tagger is measured in terms of accuracy i.e. the percentage of words, which are accurately tagged by the tagger. The accuracy has been measured using the following formula [12],

$$Accuracy = \frac{CorrectlyTaggedWords}{TotalNo.ofNumberTagged}$$

For evaluating proposed tagger, a corpus having text from special homoeopathy books, medical reports, symptoms and prescriptions. The outcome was manually appraised to mark the correct and incorrect tag assignments. 123 sentences (2319 words) collected randomly from 20060 words corpus of homoeopathy were manually appraised and are grouped into four different diseases. Only four diseases are to be taken from the complete corpus for tagging.

Table 1. Performance of Part of Speech Tagger

Corpus	*Diseases Tagged Words*		*Total Words*
	Incorrect Tag	*Correct Tag*	
Rheumatism	23	321	344
Anaemia	54	722	776
Migraine	20	110	130
Keloids	118	806	924
Total	215	1959	2174

Table 2. Accuracy of POS tagging

Diseases (from corpus)	Accuracy (%) of Correctly Tagged words
Rheumatism	93.31 %
Anaemia	93.04 %
Migraine	84.62 %
Keloids	87.23 %
Average accuracy	89.55 %

Table 1 shows the performance of part of speech tagger, sentences are collected from the manually built clinical (homoeopathy) corpus. We acquired sentences from some of the diseases like Rheumatism, Anaemia, Migraine, Keloids. Correctly tagged words from Rheumatism are 321 and incorrectly tagged words are 23. From Anaemia 722 words are correctly tagged and 54 words are incorrectly tagged. From Migraine 110 correctly tagged words and 20 incorrectly tagged words. And from Keloids 806 words correctly tagged and 118 words incorrectly tagged. Hence total tagged words

are 2174 out of which 1959 are correctly tagged and 215 are incorrectly tagged. The accuracy of POS tagging is revealed in the table 2.

From table 2. Accuracy of correctly tagged words from Rheumatism is 93.31%, Anaemia is 93.04%, Migraine is 84.62%, and Keloids is 87.23%.Total accuracy achieved by the proposed tagger is 89.55%.

7 Conclusion

The proposed Part of Speech tagger of Hindi Language to tagged Homoeopathy was developed manually. The resulting accuracy was computed to 89.55%. We use untagged Homoeopathic corpus of 20060 words, corpus is categories into different diseases. We computed correctly and incorrectly tagged words 1959 and 215 respectively. For tagging we had assembled four diseases (Rheumatism, Anaemia, Migraine, and Keloids). Sentences of each disease were autonomously tagged with accuracy 93.31%, 93.04%, 84.62%, and 87.23%, respectively, and the average percentage is computed to 89.55%. To acquire further accuracy, more data is required. In addition to that, data should be taken from homoeopathy books, patient's medical report and symptoms of different diseases. We plan to broaden the homoeopathy corpus up to 1, 50,000 words to get the better results of tagging.

References

1. Jurafsky, D., Martin, J.H.: Word classes and Part-Of-Speech Tagging. In: Speech and Language Processing, ch. 8. Prentice Hall (2000)
2. Halevi, Y.: Part of Speech Tagging. In: Seminar in Natural Language Processing and Computational Linguistics, School of Computer Science, Tel Aviv University, Israel (April 2006)
3. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics 21(4), 543–565
4. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: Brill, E., Church, K. (eds.) Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 133–142. Association for Computational Linguistics, Somerset
5. Merialdo, B.: Tagging English text with a probabilistic model. Computational Linguistics 20(2), 155–171
6. Brants, T.: TnT-a statistical part-of-speech tagger. In: Proceedings of the 6th Applied NLP Conference, ANLP-2000 (April 2000)
7. Schulze, B.M., et al.: Comparitive State-of-the-art Survey and Assessment of General Interest Tools, Technical Report DIB – I, DECIDE Project, Institute for Natural Language Processing, Stuttgart (1994)
8. Antony, P.J., Mohan, S.P., Soman, K.P.: SVM Based Part of Speech Tagger for Malayalam. In: IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 339–341 (2010)
9. Gim'enez, J., M'arquez, L.: SVMTtool: Technical manual, vol. 3 (August 2006)

10. Samuelsson, C., Voutilainen, A.: Comparing a linguistic and a stochastic tagger. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, pp. 246–253. Association for Computational Linguistics, Morristown
11. Kuba, A., Hócza, A., Csirik, J.A.: POS Tagging of Hungarian with Combined Statistical and Rule-Based Methods. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 113–120. Springer, Heidelberg (2004)
12. Kumar, D., Josan, G.S.: Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. International Journal of Computer Applications (0975-8887) 6(5), 1–9 (2010)