Word Sense Disambiguation Approaches for Indian Languages: A Survey

Misha Mittal
Research Scholar
Guru Kashi University
Talwandi Sabo

Dinesh Kumar
Associate Professor
Guru Kashi University
Talwandi Sabo

ABSTRACT

In natural languages, there are many words that have different meaning in different context. Word sense disambiguation is a method for identification of correct word sense in a specific context. Indian languages are morphologically rich in nature and hence show more inflections as compare to foreign languages. Therefore it is difficult to develop word sense disambiguation system for Indian languages. In this paper various efforts done by various researchers to develop word sense disambiguation systems for Indian languages have been discussed.

Keywords

POS tagger, word sense disambiguation, Indian languages

1. INTRODUCTION

The work on Part-of-Speech (POS) tagging has begun in the early 1960s. Word sense disambiguation means selecting the correct part of speech and hence it is also called POS tagging. It is the basic building block of many Natural Language Processing (NLP) tool. A POS tagger has many applications. Especially for Indian languages, POS tagging adds many more dimensions to Indian languages as most of them are morphologically very rich and highly inflected. Word sense disambiguation system for Indian languages have developed using linguistic rules, stochastic models or both.

India is a large multi-lingual country of diverse culture. It has many languages with written forms and over a thousand spoken languages. The Constitution of India recognizes 22 languages, spoken in different parts the country. The languages can be categorized into two major linguistic families namely Indo Aryan and Dravidian. These classes of languages have some important differences. Their ways of developing words and grammar are different. But both include a lot of Sanskrit words. In addition, both have a similar construction and phraseology that links them close together.

There is a need to develop information processing tools to facilitate human machine interaction, in Indian Languages and multi-lingual knowledge resources. A word sense disambiguation system is an integral part of any such processing tool to be developed. POS Tagging involves selecting the most likely sequences of syntactic categories for the words in a sentence. The process of POS tagging consists of three stages. These include Tokenization, Assign a tag to tokenized word and search for Ambiguous word. For disambiguation linguistic feature of the word are analyzed, it’s preceding word, its following word are analyzed.

2. WORD SENSE DISAMBIGUATION APPROACHES

Word sense disambiguation approaches can be broadly categorized in to two types i.e. supervised and unsupervised models. Further classification of these models is shown in figure 1.

2.1 Supervised Models

As shown in figure 1. There are three supervised models. These are rule based, stochastic based and neural network based. All these models need pre annotated corpus (i.e. corpus with tag associated to each word). This preannotated corpus is used for training. Training means learn information about tagset, rule set, word tag frequencies etc. the efficiency of this model depends upon the size of annotated corpus. More is the size of corpus, more will be the accuracy.

2.2 Unsupervised Models

These types of models do not required pre annotated corpus for training. These system uses advanced computational techniques like Baum-Welch Algorithm to automatically induce the tagset, transformation rules etc. this information is used to calculate the probabilistic information needed to stochastic taggers to induce the contextual rules needed by rule based system.

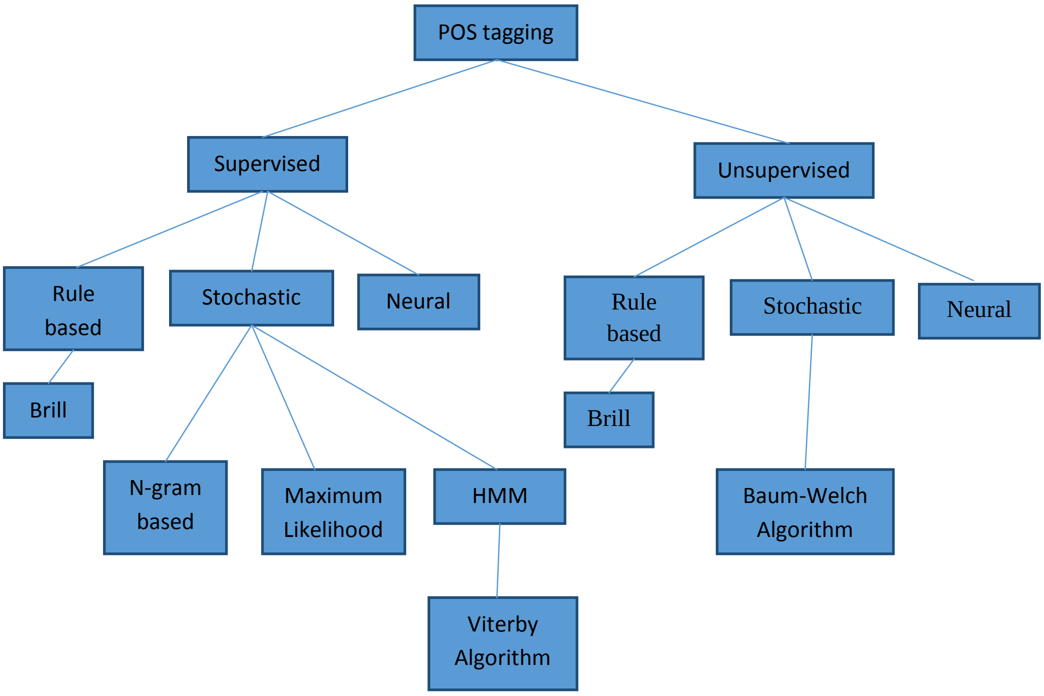

Fig 1.Classification of POS tagging models

3. WORD SENSE DISAMBIGUATION SYSTEMS FOR INDIAN LANGUAGES

Many systems have been developed by different researches for different Indian languages.

.

- **M S Gill and G SLehal in 2008** proposed a rule based Punjabi POS tagger. This POS tagger was developed as a part of Punjabi grammar checker and hence the tagset used in this POS tagger was very large and contains more than 630 tags. This POStagger used hand written rules.
- **SK Sharma and G S Lehal in 2011** proposed a HMM based POS tagger. This POS tagger uses *viterby* algorithm to implement the Hidden Markov Model.
- **Hammad Ali in 2010** proposed an unsupervised POS tagger for the Bangla language based on a Baum-Welch trained HMM approach [3]. The proposed Layered Parts of Speech Tagger is a rule based system, with four levels of layered tagging. The tagset used in the POS tagger was based on common tag set for Indian Languages and IIIT tagset guidelines. In the first level, a universal category containing 12 different categories are identified which is used to assign ambiguous basic category of a word. Followed by the first level, disambiguation rules are applied in the second level with more detail morphological information. The third and fourth levels are intended to tagging of multi word verbs and local word grouping. The proposed rule based approach shows better performance.
- **Nidhi Mishra and Amit Mishra in 2011** proposed a Part of Speech Tagging for Hindi Corpus [4]. In the proposed method, the system scans the Hindi corpus and then extracts the sentences and words from the given corpus. Also the system search the tag pattern from database and display the tag of each Hindi word like noun tag, adjective tag, number tag, verb tag etc.
- **Antony P.J, Santhanu P Mohan, Soman K.P** proposed a tagger for Malayalam was proposed [5] which is based on machine learning approach with Support Vector Machine (SVM) . There objective was to identify the ambiguities in Malayalam lexical items, and to develop a tag set appropriate for Malayalam. Finally, to built an efficient and accurate POS Tagger. The proposed tagset for Malayalam language has 29 tags where there are 5 tags for nouns, 1 tag for pronoun, 7 tags for verbs, 3 for punctuations, two for number, and 1 for each adjective, adverb, conjunction, echo, reduplication, intensifier, postposition, emphasize, determiners, complimentizer and question word. Author used SVM tool for tokenization and the desired input in column format was given to this tool. In the initial phase, 20,000 words were tagged manually. The manually tagged corpus was trained using SVM tool. This output of the tool was a dictionary with merged model and its lexicon. The remained pre-edited corpus was given to the SVM (SVMTagger, component of SVM tool) for tagging in step by step. After tagging, the displayed output

was checked manually and the tags are corrected properly. The proposed POS tagger has a tagged Malayalam corpus with size of 1, 80,000 tagged words.

- **Ekbal, A. Bandyopadhyay, S** in their work, SVM based approach was used for the task of POS tagging. To improve the accuracy of the POS tagger, a lexicon and a CRF-based NER system was used, along with the variety of contextual and word level features. The SVM based POS tagger had been developed using a corpus 72,341 word forms tagged with the 26 POS tags, defined for the Indian languages. Out of 72,341 word forms, around 15K word forms had been selected as the development set and the rest, i.e., 57,341 word forms had been used as the training set of the SVM based tagger in order to find out the best set of features for POS tagging in Bengali. The baseline model had been defined as the one where the POS tag probabilities depend only on the current word. In their model, each word in the test data was assigned the POS tag, which occurred most frequently for that word in the training data. Features for part of speech (POS) tagging in Bengali had been identified based on the different possible combination of available word and tag context. The features also included prefix and suffix for all words. A standard test set of 20K word forms was used in order to report the evaluation results of the system. The POS tagger had demonstrated the overall accuracy of 86.84% for the test set by including the unknown word handling mechanisms.
- **Nisheeth Joshi, Hemant Darbari and ItiMathur in 2013** proposed a HMM based Part of speech tagger for Hindi language. They used HMM based statistical technique to train their POS tagger for Hindi. They disambiguated correct word-tag combinations using the contextual information available in the text. They attained the accuracy of 92.13% on test data.
- **Manju et. al.** [7] proposed an HMM based tagger for Malayalam, since they did not had an annotated corpus, they used a morphological analyzer to generate the corpus which was then used for training the HMM algorithm. Another tagger for Malayalam was developed by Anthonyet. al. [7] who used Support Vector Machines (SVM). They used a SVMTool for tagging which was developed by Giménez and Màrquez [8]. For developing this tagger Anthony et. al. first proposed a tagset which they claim is suitable for Malayalam and then created an annotated corpus using this tagset. Their tagger reported 94% accuracy with their tagset.
- **H.B. Patil, A.S. Patil, B.V. Pawar** developed a Part of Speech Tagger for Marathi Language. They used Limited Training Corpora. This technique is similar to rule based technique. Here sentence taken as an input generated tokens. Once token generated apply the stemming process to remove all possible affix and reduce the word to stem. SRR used to convert stem word to root word. They developed 25 SRR rule.
- **For Bengali, Dandapatet. al.[9]** studied the possibility of developing a tagger using HMM and Maximum Entropy (ME) models. They too used a morphological analyzer for compensating theshortage of annotated corpus. With these two modes they implemented a supervised tagger and asemi-supervised tagger and reported an accuracy of around 88% for the two approaches. Ekbaland Bandyopadhyay[10] annotated news corpus and developed an SVM based tagger. They reported an accuracy of 86.84% for their tagger.
- **Ekbalet. al.** [11] also developed a Conditional Random Fields(CRF) based tagger. For training the tagger they used the information of prefix and suffix of Bengali words along with normal word/tags and reported an accuracy of 90.3%. For Tamil, Selvam and Natarajan[12] proposed a POS tagger which used a rule based morphological analyzer to annotate the corpora which was used to train the tagger. They used the Tamil version of the Bible for annotation of POS tagged corpus and reported an accuracy of 85.56%.
- **Dhanalakshmiet. al.**[13] proposed an SVM based tagger using linear programming and developed their own POS tagset for Tamil which has 32 tags. They used this tagset to annotate their corpus and then trained their model and reported an accuracy of 95.63%. Dhanalakshmiet.
- al.[14] also proposed another tagger where they used machine learning techniques to extract linguistic information which was then used to train the tagger based on SVM approach. They used their own 32 tags tagset for annotating the corpus and reported an accuracy of 95.64%.
- **ForMarathi, Singh et al** [15] proposed a POS tagger using trigram method. They used a postagsetproposed by Bharti et al [16] which had 24 tags. They showed an accuracy of 91.63%.

A summary of commonly developed POS tagging systems is provided in table 1.

Table 1. Word sense disambiguation systems developed for Indian Languages

Sr. No.	Language	Technique used	Year
1	Punjabi	Rule based	2008
2	Punjabi	HMM based	2011

3	Bangla	Baum-Welch trained HMM approach	2010
4	Hindi	Pattern matching	2011
5	Malayalam	Support Vector Machine and HMM	2010 and 2009
6	Hindi	HMM	2013
7	Bengali	CRF	2007
8	Tamil	CRF SVM	2009
9	Marathi	Rule based	2014
10	Telugu	Machine learning	
11	Sanskrit	Treebank based	2012

4. CONCLUSION

At last this is concluded that part of speech tagging is one of the most important activities in the natural language processing. The accuracy of most of natural language applications largely depends upon the accuracy of part of speech tagger. Various approached have been used by different authors to improve the accuracy of part of speech tagger. Even for a single language more than one approach has been tried to increase the efficiency of the part of speech tagger.

REFERENCES

[1] 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) Volume6–No.5, September, 2010, www.ijcaonline.org/ volume6/number5 /pxc3871409 .pdf..

[2] Vijayalaxmi .F. Patil (2010), “Designing POS Tagset for Kannada, Linguistic Data Consortium for Indian Languages (LDC-IL), Organized by Central Institute of Indian Languages, Department of Higher Education Ministry of Human Resource Development, Government of India, March 2010..

[3] Hammad Ali (2010), “An Unsupervised Parts-of-Speech Tagger for the Bangla language”, Department of Computer Science, University of British Columbia. 2010.

[4] Nidhi Mishra Amit Mishra (2011), “Part of Speech Tagging for Hindi Corpus”, International Conference on Communication Systems and Network Technologies.

[5] Antony P.J, Santhanu P Mohan, Soman K.P,”SVM Based Part of Speech Tagger for Malayalam”, IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 339-341, 2010.

[6] Ekbal, A. Bandyopadhyay, S., “Part of Speech Tagging in Bengali Using Support Vector Machine”, ICIT-08, IEEE International Conference on Information Technology, pp. 106-111, 2008.

[7] Manju K., Soumya S., Sumam, M. I., (2009) “Development of a POS Tagger for Malayalam – AnExperience”. In: International Conference on Advances in Recent Technologies in Communication and Computing, pp.709-713.

[8] Jesus Giménez and LlúisMàrquez., (2006) “SVMTool. Technical manual v1.3”.

[9] Dandapat, S., Sarkar, S., Basu, A., (2007) “Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario”. In: Association for Computational Linguistic, pp 221-224.

[10] Ekbal, A., Bandyopadhyay, S., (2007) “Lexicon Development and POS tagging using a Tagged Bengali News Corpus”. In: FLAIRS-2007, Florida, pp 261-263.

[11] Ekbal, A., Haque, R., Bandyopadhyay, S., (2007) “Bengali Part of Speech Tagging using Conditional Random Field”. In: 7th International Symposium of Natural Language Processing(SNLP-2007),Thailand Pattaya, 13-15 December 2007, pp.131-136.

[12] Selvam, M., Natarajan, A.M., (2009) “Improvement of Rule Based Morphological Analysis and POSTagging in Tamil Language via Projection and Induction Techniques”. International Journal of Computers, 3(4).

[13] Dhanalakshmi, V., Kumar, A., Shivapratap, G, Soman, K.P., Rajendran, S, (2009) “Tamil POSTagging using Linear Programming”. International Journal of Recent Trends in Engineering, 1(2).

[14] Dhanalakshmi V, Anandkumar M, Rajendran S, Soman K P., (2009) ”POS Tagger and Chunker forTamil Language”. Proceedings of Tamil Internet Conference 2009.

[15] Singh, J., Joshi, N., Mathur I., (2013) “Part of Speech Tagging of Marathi Text Using TrigramMethod”, International Journal of Advanced Information Technology, pp 35-41, Vol 3. No. 2.

[16] Bharati, A., Sharma, D.M., Bai, L., Sangal, R., (2006) “AnnCorra: Annotating Corpora Guidelines forPOS and Chunk Annotation for Indian Languages”, http://ltrc.iiit.ac.in/tr031/posguidelines.pdf