A Brief Survey on Part of Speech Tagging Techniques Used in Punjabi Language and Other Regional Languages

[1]Rabia Sachdeva, [2]Abhilasha,
[1]GZS, Bathinda, [2]Associate professor (CSE), Giani Zail Singh Punjab Technical University Bathinda
[1]rabia.sachdeva01@gmail.com
[2]abd_jain@rediffmail.com

ABSTRACT-*Parts-of-speech (POS) tagging is the basic building block of any Natural Language Processing (NLP) tool. A POS tagger has many applications. POS tagger represents structure and forms of words and phrases of a particular language. Therefore, these can be used in different research fields in NLP. Taggers can be developed using linguistic rules, stochastic models or both .This paper presents a review about different POS taggers developed for different Indian Language.*

KEYWORDS-*Parts-of-speech tagger, Indian languages, HMM technique.*

INTRODUCTION

Part of speech (POS) tagging is process of analysis the text and mark each word in text according to its part of speech based on its definition as well as relationship with the adjacent word. It provides useful information to a language processor in pipeline. Therefore it may be viewed as the basic step that is to be performed in every NLP task.

Commonly used parts of speech in any language are: nouns, verb, Adjective, Preposition, Article, Pronoun, Conjunction, Adverb and Interjection and their sub-categories. E.g. NN for Noun, NNS for Singular Common noun etc. In case of morphologically rich languages tags are commonly expressed as '*NNMPD* for Category=Noun, Gender = Masculine, Number = Plural, Case = Direct. They provide different information systems in different applications of NLP. E.g. in summarization these parts of speech tags gives the information of proper noun, in grammar checking these tags gives the information of word class, gender, number and case etc. The main purpose of POS tagging is to select the appropriate tag for a word that can exist in more than one form in different contexts.

When a word can act in more than one form in different contexts due to this ambiguity arises. This is a common problem in almost all the languages. The amount of ambiguity depends upon various factors like the type of language.

More ambiguities occur in Indian languages because Indian languages are more reflected as compare to other languages. One more factor which decides the amount of ambiguity is the tag set used for assigning the tags. If we will use the fined grained tag set then there will be more ambiguities. So the length of the tag directly related with the ambiguity. POS tagging is a type of classification problem so many algorithms have been proposed to solve it. There are basically two approaches used to solve this problem. These are rule based approach and other is statistical based approach. Initially rule based approach has been tries for almost all languages. This approach was used because:

1. It is easy to implement.
2. No need of annotated corpus.
3. Rule can be on and off as per requirement.
4. More rules can be added and existing rules can be modified.

5. Rules can be inherited from similar languages.

The disadvantages of this approach are:

1. Thorough Language knowledge is required.
2. Exhaustive set of rules have to be developed.
3. This method fails in case of unknown words.

Another approach used for part of speech tagging is statistical based approach. All the limitations of rule based approach are overcome by statistical based approach. E.g.

1. There is no need of thorough linguistic knowledge.
2. Unknown words will be given the tag based upon probability.
3. A system can be trained by using a small corpus.

Statistical methods are considered to be better than rule based methods because these methods are probabilistic based. Another advantage of these methods is

that we can use feature values to increase the accuracy as observed in case of conditional random fields.

CURRENT WORK IN INDIAN LANGUAGES

As India is a multilingual country. Different languages are spoken in different region of India. There are 23 state languages recognized by constitution of India. Every state is working on his language. POS tagger has been developed for most of the Indian languages using one or combination of more than one of the following techniques:

1. Rule based technique
2. Statistical based technique.

The commonly used statistical techniques are:

1. HMM (Hidden Markov Model)
2. CRF(Conditional Random Field)
3. SVM (Support Vector Machine)
4. ME (Maximum Entropy based technique)
5. N-gram model based technique.
6. Hybrid model

One or more than one of the above mention techniques have been applied for most of the Indian languages. Accuracy of these techniques varies from language to language. SVM and CRF are feature based techniques, so they give better result for morphologically rich languages like Indian languages. The basic resource for all the statistical approaches is annotated corpus.

Rule based system:

In 2006, Sreeganesh [28] implemented a rule based POS tagger on telugu language. The whole process takes place in two steps. In the first step the input text is analysed by a telugu morphological analyser and all possible tags are assigned to each words. In the second step 524 morpho-syntactic rules are used for disambiguation. Another rule based POS tagger was developed for Punjabi language. Mandeep Singh Gill, Gurpreet

Singh Lehal (2008) [17] has used rule-based approaches for part-of-speech tagging.

HMM based system:

A simple HMM based POS tagger for Hindi language was designed by Manish Shrivastava & Pushpak Bhattacharyya [18]. This pos tagger was designed by using the morphological richness of the language and achieved an accuracy of 93.12%. Another POS tagger for malyalam language was developed by Manju K et. al [19]. It also works in two stages. In the first stage an annotated corpus was developed by using morphological analyzer. This annotated tagger was used to design the HMM based tagger. One more POS tagger for assamese was developed by Navanath Saharia et.al

[20] using the HMM model with viterbi algorithm. An accuracy of this POS tagger was reported as 87%. Another efforts were done by Sanjeev Kumar Sharma and G S Lehal (2011)[23] to develop an HMM based Part of speech tagger for Punjabi language. They also implement the HMM using viterby algorithm. Ekbal, S. Mondal and S. Bandyopadhyay (2007) [9] also developed a HMM based POS tagger for Bengali language.

SVM based system:

Tagset Ekbal and S. Bandyopadhyay (2008) [12] developed a SVM based POS taggers for Bengali language. They used a tag set of 26 tags and the accuracy was reported to be 86.84%. Two SVM based POS tagger were developed for Tamil language. One was developed by V.Dhanalakshmi, M Anandkumar, Vijaya M.S, Loganathan R, Soman K.P, Rjendran S (2008) [31]. They used a freely available tool i.e. SVM Tool for training the corpus an dthey obtained an accuracy of 94.12%. Other efforts were again done by Dhanalakshmi et. al.[32]. They used SVM methodology based on Linear programming. They were successful to improve the accuracy to 95.63%. Other efforts using this technology were done by Sindhiya Binulal et. al [15] who had applied SVM Tool for POS tagging of Telugu language. They used a tag set of 10 tags and achieved an accuracy of around 95%. Malayalam was the Indian language for which SVM technique was used for development of POS tagger Antony P.J et. al [4] applied SVM approach for POS tagging of Malayalam language. They used a tag set of 29 tags. The result was more accurate as compared to earlier work. The performance increased to around 94% by increasing the no. of words in the training set.

ME based system:

This technique has been used for POS tagging of Hindi language. Maximum Entropy Markov Model was developed by Aniket Dalal et. al [2] Developed a Maximum Entropy Markov Model. In this model they used main POS tagging features like context based features, word features, dictionary features and corpus-based features. Other efforts were done for Bengali language. Ekbal, R. Haque and S. Bandyopadhyay (2008) [11], developed Maximum Entropy based tagger. An accuracy of 88.2% was reported by this tagger.

CRF based system:

This technique was first applied for POS tagging of Hindi language. Ravindran et. al. [21] and Himanshu et. al.[16] used CRF for POS tagging and chunking. They achieved an accuracy of 89.69% for POS tagging and 90.89% for phrase chunking. Other Indian languages on which this CRF technique has been applied are Bengali [10] and Manipuri [30].

Neural network based system:

The only POS tagger developed by using this technique was for Hindi language. This tagger was developed by Ankur Parikh [3].neurons were trained for tagging. This system perform better on small data and need less training as compare to other systems A recent work is also going on the Punjabi language.

Hybrid systems:

Arulmozhi.P, L Sobha (2006) used a combination of rule based and HMM based technique for development of Tamil POS tagger. Rama Sree, R.J., P Kusuma Kumari (2007)[22] used three Telugu taggers namely (i) Rule-based tagger, (ii) Brill Tagger and (iii) Maximum Entropy tagger with accuracies of 98.016%, 92.146%, and 87.81% respectively for development of hybrid system. Another hybrid POS tagger was developed for Gujarati language by Chirag Patel and Karthik Gali [8] . they used a combination of rule based method and CRF and hence took the advantage of rule based and statistical based system. They were succeeded to achieve an accuracy of 92%.

POS tagging of Punjabi Language

Punjabi (or Panjabi) language is a member

of the Indo-Aryan family of languages, also known as Indic languages. Hindi, Bengali, Gujarati, and Marathi etc are the other members of this family. Mandeep Singh Gill, Gurpreet Singh Lehal (2008) [17] has developed a Grammer Checking System for Punjabi. They provided description about the grammar checking system developed for detecting various grammatical errors in Punjabi texts. This system applies a full-form lexicon for morphological analysis, and uses rule-based approaches for part-of-speech tagging and phrase chunking. The system adopts a novel approach of performing agreement checks at phrase and clause levels using the grammatical information exhibited by POS tags in the form of feature value pairs. The system can observe and propose rectifications for a number of grammatical errors, resulting from the deficiency of agreement, order of words in several phrases etc, in literary style Punjabi texts. This grammar checking system is the first such system reported for Indian languages. Sanjeev Kumar Sharma, G S Lehal (2011)[23] has developed a HMM based Part of speech tagger for Punjabi language. They implement the HMM using viterby algorithm. They used previously developed tagset of 630 tags. They also tried a hybrid approach that is combination of rule based system and statistical approach in which the output of rule based system was fed to the statistical based system. This gives further improvement on the accuracy of the POS tagger.

PART OF SPEECH TAGGING IN PUNJABI LANGUAGE:

Ambiguous words are the main problem in part of speech tagging. Many words may have more than one tag. Sometimes a word has same POS but have different meaning in different context. To solve this problem we consider the sentence instead of taking single word.

For example-

ਗੰਭੀਰ_AJU ਸੋਚ_NNFSD ਤੇ_PTUE

ਦ੍ਰਿੜ੍ਹ_AJU ਇਰਾਦੇ_NNMSO ਨਾਲ਼_PPU

ਉਹ_PNDBSD|PNDBPD|IJ ਅੱਗੇ_AVIBSD

ਵਧਦੀ_VBMAFSXXXINDA

ਗਈ_VBMAFSXXPINIA ।_Sentence

The same word ‘ਉਹ’ is given more than one label in a same sentence. In the first case it is termed as a singular pronoun. In the second case it is termed as a plural pronoun and in the third case it may be tagged as interjection. Since word ਉਹ occur in between the sentence and also the word next to it is not a noun so it may be a pronoun and not an interjection. Now the type of pronoun that is singular or plural depends upon the previous words of the sentence. POS Tagging tries to correctly identify a POS of a word by looking at the context (surrounding words) in a sentence.

EXISTING PUNJABI POS TAGSET

Sr. No.	Word Class	Tagset
1.	Noun	NN<a><b><c> - Noun<a><b><c> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular) or P (Plural) • <c> can be D (Direct), O (Oblique), V (Vocative), A (Ablative), L (Locative),or I (Instrumental): e.g. NNMSD, NNFPI, NNBSL.

2.	Personal Pronoun	PNP<a><b><c><d> - Pronoun Personal<a><b><c><d> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular) or P (Plural) • <c> can be F (First), S (Second), or T (third) • <d> can be D (Direct), O (Oblique), A (Ablative), T (Dative) or V (Vocative): e.g. PNPMSFD, PNPFPTO, PNPBSFT
3.	Reflexive Pronoun	PNR<a><b><c> - Pronoun Reflexive<a><b><c> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular) or P (Plural) • <c> can be D (direct), O (Oblique), or V (Vocative): e.g. PNRMSD, PNRFPV, PNRBPO
4.	Demonstrative Pronoun	PND<a><b><c> - Pronoun Demonstrative<a><b><c> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular) or P (Plural) • <c> can be D (Direct), O (Oblique), L (Locative), A (Ablative), T (Dative), or I (Instrumental): e.g. PNDMSD, PNDFPO, PNDBPI
5.	Indefinite Pronoun	• PNIBSD – Pronoun Indefinite Both Singular Direct • PNIBSO – Pronoun Indefinite Both Singular Oblique • PNIBPD – Pronoun Indefinite Both Plural Direct • PNIBPO – Pronoun Indefinite Both Plural Oblique • PNIBPL – Pronoun Indefinite Both Plural Locative • PNIBPI – Pronoun Indefinite Both Plural Instrumental
6.	Relative Pronoun	PNE<a><b><c> - Pronoun Relative<a><b><c> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular) or P (Plural) • <c> can be D (Direct), O (Oblique), or I (Instrumental) : e.g. PNEMSD, PNEFPO, PNEBPI
7.	Interrogative Pronoun	PNN<a><b><c> - Pronoun Interrogative<a><b><c> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular) or P (Plural) • <c> can be D (Direct), O (Oblique), A (Ablative), or I (Instrumental) : e.g. PNNMSD, PNNFPO, PNNBSA
8.	Inflected Adjective	AJI<a><b><c> - Adjective Inflected<a><b><c> <a>, <b>, and <c> will have values same as the noun word class.

9.	Uninflected Adjective	AJU – Adjective Uninflected
10.	Cardinal	• CDPD – Cardinal Plural Direct • CDPO – Cardinal Plural Oblique • CDPL – Cardinal Plural Locative
11.	Ordinal	• ODMSD – Ordinal Masculine Singular Direct • ODMSO – Ordinal Masculine Singular Oblique • ODFSD – Ordinal Feminine Singular Direct • ODFSO – Ordinal Feminine Singular Direct
12.	Main Verb	VBM<a><b><c><d><e><f><g><h> - Verb Main<a><b><c><d><e><f><g><h> • <a> can be M (Masculine), F (Feminine), or X (No Inflection) • <b> can be S (Singular), P (Plural), or X (No Inflection) • <c> can be F (First), S (Second), T (Third), or X (No Inflection) • <d> can be 1 (Present), 2 (Past), 3 (Future), or X (No Inflection) • <e> can be P (Perfect) or X (No Inflection) • <f> can be T (Transitive), I (intransitive), or B (Both) • <g> can be N (None), S (Single Causal), or D (Double Causal) • <h> will be inflectional class value e.g. EGA, IA, DA etc. This serves the purpose of 'mood' and helps in phrase chunking and checking phrase concordance, in later stages of grammar checking: e.g. VBMMSF3XTNEGA, VBMMSXXPTNIA, VBMMPS3XTNEGA.
13.	Auxiliary Verb	VBA<a><b><c><d> - Verb Auxiliary<a><b><c><d> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular) or P (Plural) • <c> can be A (All), F (First), S (Second), or T (Third) • <d> can be 1 (Present) or 2 (Past): e.g. VBAMSA1, VBFPA2, VBABST1.
14.	Inflected Adverb	AVI<a><b><c> - Adverb Inflected<a><b><c> <a>, <b>, and <c> will have values same as noun word class.
15.	Uninflected Adverb	AVU – Adverb Uninflected
16.	Inflected Postposition	PPI<a><b><c> - Postposition Inflected<a><b><c> <a>, <b>, and <c> will have values same as noun word class.
17.	Uninflected Postposition	PPU – Postposition Uninflected
18.	Conjunction	• CJC – Conjunction Coordinating • CJS – Conjunction Subordinating

19.	Interjection	IJ – Interjection
20.	Particle	• PTUE – Particle Uninflected Emphatic • PTUN – Particle Uninflected Negative • PTUH – Particle Uninflected Honorific
21.	Vocative Particle	PTV<a><b> - Particle Vocative<a><b> • <a> can be M (Masculine), F (Feminine), or B (Both) • <b> can be S (Singular), P (Plural), or B (Both) : e.g. PTVMS, PTVFP, PTVBB
22.	Verb-Part	VBP – Verb Part

CONCLUSION

Development of a high accuracy POS tagger is an active research area in NLP. The bottleneck to POS tagging of Indian languages is the non-availability of lexical resources. In addition, adoption of common tagset by researchers would facilitate reusability and interoperability of annotated corpora. We have presented in this paper a detailed study of the POS taggers developed for different Indian languages. But there exist many other languages of the country, for which hardly any attempts towards building a POS tagger have started. Also we observed that there is not much work has been done on Punjabi language due to lack of annotated corpora. Also the tagset used in the existing system is very large. So we can apply some statistical techniques with reduces tagset to develop the POS tagger.

REFERENCES

[1] Ahmed, Raju S.B, Chandrasekhar Pammi V. S., Prasad M.K (2002), "Application of multilayer perceptron network for tagging parts-of-speech", Proceedings of the Language Engineering Conference, IEEE.

[2] Aniket Dalal, Kumar Nagaraj, Sawant Uma, Shelke Sandeep (2006), "Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach" Proceedings of the NLPAI MLcontest workshop, National Workshop on Artificial Intelligence.

[3] Ankur Parikh (2009), "Part-Of-Speech Tagging using Neural network", Proceedings of ICON-2009: 7th International Conference on Natural Language Processing.

[4] Antony P.J, Mohan S. P., Soman K.P (2010), "SVM Based Part of Speech Tagger for Malayalam", Proceedings of 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, IEEE.

[5] Anupam Basu, Ray, Ranjan Pradipta, Harish V. and Sarkar Sudeshna(2003), "Part of speech tagging and local word grouping techniques for natural language parsing in Hindi", Proceedings of the International Conference on Natural Language Processing (ICON 2003).

[6] Arulmozhi.P, L Sobha (2006) "A Hybrid POS Tagger for a Relatively Free Word Order Language", Proceedings of MSPIL-2006, Indian Institute of Technology, Bombay.

[7] Avinesh PVS and Gali Karthik (2007), "Part-of-speech tagging and chunking using conditional random fields and transformation based learning", Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL), pp. 21–24.

[8] Chirag Patel and Gali Karthik (2008), "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, Hyderabad, India, pp. 117–122.

[9] Ekbal, S. Mondal and S. Bandyopadhyay (2007). POS Tagging using HMM and Rule-based Chunking. In Proceedings of the Workshop on Shallow Parsing in South Asian Languages, International Joint Conference on Artificial Intelligence (IJCAI 2007), 6-12 January 2007, Hyderabad, India, PP. 25-28.

[10] Ekbal, R. Haque and S. Bandyopadhyay (2007), "Bengali Part of Speech Tagging using Conditional Random Field", Proceedings of the 7th International Symposium on Natural Language Processing (SNLP-07), Thailand, pp.131-136.

[11] Ekbal, R. Haque and S. Bandyopadhyay (2008), "Maximum Entropy Based Bengali Part of Speech Tagging", Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal, Vol. (33), pp. 67-78.

[12] Ekbal and S. Bandyopadhyay (2008), "Part of Speech Tagging in Bengali using Support Vector

Machine", Proceedings of the International Conference on Information Technology (ICIT 2008), pp.106-111, IEEE.

[13] Ekbal , M. Hasanuzzaman and S. Bandyopadhyay (2009), "Voted Approach for Part of Speech Tagging in Bengali", Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC-09), December 3-5, Hong Kong, pp. 120-129.

[14] Ganesan M (2007), "Morph and POS Tagger for Tamil" (Software) Annamalai University, Annamalai Nagar.

[15] G.Sindhiya Binulal, Goud P. A, K.P.Soman(2009), "A SVM based approach to Telugu Parts Of Speech Tagging using SVMTool", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[16] Himanshu Agrawal, Mani Anirudh (2006), "Part Of Speech Tagging and Chunking Using Conditional Random Fields" Proceedings of the NLPAI MLcontest workshop, National Workshop on Artificial Intelligence.

[17] Mandeep Singh Gill, Lehal G.S. (2008) "Grammer Checking System for Punjabi" Coling 2008:companion volume Posters and Demonstrations pages 149–152 Manchester.

[18] Manish Shrivastava, Bhattacharyya Pushpak (2008), "Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge", Proceedings of ICON-2008: 6th International Conference on Natural Language Processing.

[19] Manju k,S Soumya , Idicula S.M. (2009), "Development of A Pos Tagger for Malayalam-An Experience", Proceedings of 2009 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE .

[20] Navanath Saharia, Das Dhrubajyoti, Sharma Utpal, Kalita Jugal (2009), "Part of Speech Tagger for Assamese Text", Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore, pp. 33–36.

[21] Pranjal Awasthi, Rao Delip, Ravindran Balaraman (2006), "Part Of Speech Tagging and Chunking with HMM and CRF", Proceedings of the NLPAI MLcontest workshop, National Workshop on Artificial Intelligence.

[22] Rama Sree, R.J., P Kusuma Kumari (2007), "Combining POS Taggers for improved Accuracy to create Telugu annotated texts for Information Retrieval", Tirupati.

[23] Sanjeev Kumar Sharma, Lehal G.S. (2011) " improving Existing Punjabi POS tagger using Hidden Markov Model" Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on (Volume:2) pages 697-701.

[24] Sankaran Baskaran (2006), "Hindi POS tagging and Chunking", Proceedings of the NLPAI MLcontest workshop, National Workshop on Artificial Intelligence.

[25] Smriti Singh, Shrivastava Manish, Agrawal Nitin and Bhattacharya Pushpak (2005), "Harnessing morphological analysis in pos tagging task", Proceedings of the International Conference on Natural Language Processing (ICON 2005).

[26] Smriti Singh, Gupta Kuhoo, Shrivastava Manish, and Bhattacharyya Pushpak (2006),"Morphological richness offsets resource demand – experiences in constructing a pos tagger for Hindi", Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, pp. 779–786.

[27] S. Lakshmana Pandian and Geetha T.V.(2008), "Morpheme based Language Model for Tamil Part-of-

Speech Tagging", Research journal on Computer science and computer engineering with applications, July-Dec 2008, pp. 19-25.

[28] T.Sreeganesh(2006), "Telugu Parts of Speech Tagging in WSD", Language of India, Vol 6: 8 August 2006.

[29] Thoudam Doren Singh, Bandyopadhyay Sivaji (2008), "Morphology Driven Manipuri POS Tagger", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, Hyderabad, India, pp. 91–98.

[30] Thoudam Doren Singh, Bandyopadhyay Sivaji (2008), "Manipuri POS Tagging using CRF and SVM: A Language Independent Approach", Proceedings of ICON-2008: 6th International Conference on Natural Language Processing.

[31] V.Dhanalakshmi, M Anandkumar, Vijaya M.S, Loganathan R, Soman K.P, Rjendran S (2008), "Tamil Part-of-Speech tagger based on SVMTool", Proceedings of the COLIPS International Conference on Asian Language Processing 2008 (IALP), Chiang Mai, Thailand.

[32] V. Dhanalakshmi, M Anandkumar, Shivapratap G, Soman K.P, Rajendran S (2009) "Tamil POS Tagging using Linear Programming", International Journal of Recent Trends in Engineering, 1(2) pp.166-169.

[33] http://tdil.mit.gov.in/