Part of Speech Tagging of Punjabi Language using N Gram Model

Sumeer Mittal
Adhesh college of Engineering, Faridkot

Mr Navdeep Singh Sethi
Lecturer, Adhesh Institute of Engineering and Technology, Faridkot,

Sanjeev Kumar Sharma
Assistant Professor
DAV University
Jalandhar

ABSTRACT
POS tagger is the process of assigning a correct tag to each word of the sentence. We attempted to improve the accuracy of existing Punjabi POS tagger. This POS tagger lacks in resolving the ambiguity of a no of words as it uses only hand written Rules. A Bi-gram Model has been used to solve the part of speech tagging problem. An annotated corpus was used for training and estimating of bi gram probabilities.

General terms
POS tagging, NLP

Keywords
POS tagger, bi-gram, n-gram, Punjabi tag set

1. INTRODUCTION
Part-of-speech (POS) Tagging is a process assigning correct POS tag to each word in a sentence from a given set of tags. It is basic activities performed in all natural language processing application such as speech recognition, information extraction, machine translation, grammar checking and word sense disambiguation etc. This paper explores part-of-speech tagging for the Punjabi language (A member of the Modern Indo-Aryan family of languages) using N gram technique. There are many approaches for development of POS taggers. Most commonly used techniques are rule based, statistical based and neural network based. In the rule based approach, rule is developed by linguistic to define precisely how and where to assign the various POS tags. This approach has already been used to develop the POS tagger for Punjabi language. In the statistical approach, statistical language model are built, refined and used to POS tag the input text automatically. Most commonly used statistical approaches are HMM (Hidden Markov Model) based approach, SVM (Support vector machine) based, CRF (Conditional Random Field) based and ME (Maximum Entropy based approach). One of the robust approaches in statistical models is the use of N Gram model.

2. OVERVIEW OF PUNJABI LANGUAGE
Punjabi language is a member of the Indo-Aryan family of languages, also known as Indic languages. Other members of this family are Hindi, Bengali, Gujarati, and Marathi etc. Indo-Aryan languages form a subgroup of the Indo-Iranian group of languages, which in turn belongs to Indo-European family of languages. Punjabi is spoken in India, Pakistan, USA, Canada, England, and other countries with Punjabi Immigrants. It is the official language of the state of Punjab in India. Punjabi is written in "Gurmukhi" script in eastern Punjab (India), and in "Shahmukhi" script in western Punjab (Pakistan).

3. PROBLEMS OF PART OF SPEECH TAGGING
Ambiguous words are the main problem in part of speech tagging. There may be many words which can have more than one tag. Sometimes it happens that a word has same POS but have different meaning in different context. To solve this problem we consider the context 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.

4. PREVIOUSWORK ON INDIAN LANGUAGE POS TAGGING
Smriti Singh in **2010** proposed a POS tagging methodology which can be used by languages having lack of resources [1]. The POS tagger was built based on hand-crafted morphology rules and does not involve any sort of learning or disambiguation process. The system makes use of locally annotated modestly-sized corpora of 15,562 words, exhaustive morphological analysis backed by high-coverage lexicon and a decision tree based learning algorithm (CN2). The system uses Lexicon lookup for identifying the other POS categories. The performance of the system was evaluated by a 4-fold cross validation over the corpora and found 93.45% accuracy. **Vijayalaxmi .F. Patil** in **2010** developed a POS tag set for Kannada language. It used 39 tags [2]. This tag set was developed by considering the morphological as well as syntactic and semantic features of the Kannada language. **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 tag set used in the POS tagger was based on common tag set for Indian Languages and IIIT tag set 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. Jyoti Singh et all in 2013 proposed a POS tagging system for Marathi language using N-Gram method. They used trigram model for POS tagging. They obtained an accuracy of 91.6%.

5. EXISTING SYSTEM IN PUNJABI LANGUAGE:

A rule based part-of-speech tagging approach was used for Punjabi, which is further used in grammar checking system for Punjabi [14]. This is the only tagger available for Punjabi Language. A part-of-speech tagging scheme based entirely on the grammatical categories taking part in various kinds of agreement in Punjabi sentences has been proposed and applied successfully for the grammar checking of
Punjabi [14]. This tagger uses handwritten linguistic rules to disambiguate the part of-speech information, which is possible for a given word, based on the context information. A tag set for use in this part-of-speech tagger has also been devised to incorporate all the grammatical properties that will be helpful in the later stages of grammar checking based on these tags. This part-of-speech tagger can be used for rapid development of annotated corpora for Punjabi. There are around 630 tags in this fine-grained tag set. This tag set includes all the tags for the various word classes, word specific tags, and tags for punctuations. During tagging process with proposed tagger, 503 tags out of proposed 630 tags were found in 8-million words corpus of Punjabi, which was collected from online sources. For disambiguation of POS tags rule-based approach was used. A database was designed to store the rules, which is used by rule based disambiguation approach. The texts with disambiguated POS tags are than passed for marking verbal operators. Four operator categories have been established to make the structure of verb phrase more understandable. During this step the verbal operators are marked based on their position in the verb phrase and the forms of their proceeding words. A separate database was maintained for marking verbal operator. A HMM based model was used by Sanjeev Kumar Sharma and Dr G.S lehal in 2011 to develop a Part of speech tagger for Punjabi language. 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.

6. OUR APPROACH

In this paper we are describing Bigram Model for Punjabi POS tagger. Our main aim is to perform POS Tagging to determine the most likely tag for a word, given the previous and next tags. For Bigrams, the probability of a sequence is just the product of conditional probabilities of its Bigrams. So if t1, t2 …tn are tag sequence and w1, w2…wn are corresponding word sequence then the following equation explains this fact:-
$P(t_i |w_i) = P(w_i |t_i). P(t_i |t_{i+1})$

Where ti denotes the tag sequence and wi denotes the word sequences. P (w |t) is the probability of current word given current tag. Here, $P(t_i |t_{i-1})$is the probability of a current tag given the previous tag. This provides the transition between the tags and helps capture the context of the sentence. These probabilities are computed by following equation. P(ti/ti-1) = count of (ti,ti-1)/ count of (t_i , t_{i+1}). Each tag transition probability is computed by calculating the frequency count of two tags which come together in the corpus divided by the frequency count of the previous two tags coming in the corpus.

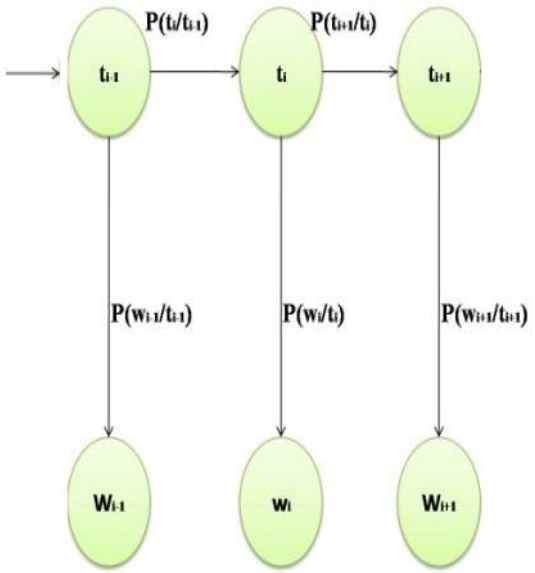

Figure 1. Working of bigram model

7. POS TAGSET

Depending on some general principle of tag set design strategy, a number of POS tag sets have been developed by different organizations based. For POS annotation of texts in Marathi, we have used tag set proposed by TDIL (Technical Development of Indian Languages). Table shows brief description of IL (Indian languages) especially for Punjabi language. POS Tag set.

Table 1. POS tag set for Punjabi Proposed by Tdil

S.No	Category	Label	Annotated convention	Example
1	Noun	N	N	ਘਰ ਕਹਾਣੀ ਸੜ਼
1.1	Common	NN	N__NN	ਘਰ ਕਹਾਣੀ ਸੜ਼
1.2	Proper	NNP	N__NNP	ਹਰਿਵੰਦਰ
1.4	Nloc	NST	N__NST	ਥੱਲੇ ਅੱਗੇ
2	Pronoun	PR	PR	ਤੂੰ ਉਹ ਇਹ ਜੋ
2.1	Personal	PRP	PR__PRP	ਤੂੰ ਉਹ
2.2	Reflexive	PRF	PR__PRF	ਆਪਣਾ

				ਆਪ ਖੁਦ
2.3	Relative	PRL	PR__PRL	ਜੋ
2.4	Reciprocal	PRC	PR__PRC	ਆਪਸ
2.5	Wh-word	PRQ	PR__PRQ	ਕੌਣ
2.6	Indefinite	PRI	PR_PRI	ਕੋਈ
3	Demonstrative	DM	DM	ਉਹ ਜੋ ਇਹ
3.1	Deictic	DMD	DM__DMD	ਇਹ ਉਹ
3.2	Relative	DMR	DM__DMR	ਜੋ
3.3	Wh-word	DMQ	DM__DMQ	ਕੌਣ
3.4	indefinite	DMI	DM_DMI	ਕੋਈ
4	Verb	V	V	ਆਇਆ ਜਾ ਕਰਦਾ ਰਿਹੰਦਾ
4.1	Main	VM	V__VM	ਆਇਆ ਜਾ ਕਰਦਾ ਮਾਰ□ਗਾ ਰਿਹੰਦਾ
4.1.2	Non-finite	VNF	V__VM__VNF	ਆਇਆਂ ਕਿਰਆਂ
4.1.3	Infinitive	VINF	V__VM__VINF	ਜਾਣ ਖਾਣ ਪੀਣ ਮਰਨ
4.1.4	Gerund	VNG	V__VM__VNG	ਹੈ ਸੀ ਸਿਕਆ ਹੋਇਆ
4.2	Auxiliary	VAUX	V__VAUX	ਸੋਹਣਾ ਚੰਗਾ ਮਾੜਾ
5	Adjective		JJ	ਹੌਲ਼ੀ ਕਾਹਲੀ
6	Adverb		RB	ਨਾਲ
7	Postposition		PSP	ਅਤੇ ਅਗਰ
8	Conjunction	CC	CC	ਅਤੇ
8.1	Co-ordinator	CCD	CC__CCD	ਜੋ
8.2	Subordinator	CCS	CC__CCS	ਵੀ ਹੀ
9	Particles	RP	RP	ਵੀ ਹੀ
9.1	Default	RPD	RP__RPD	
9.2	Classifier	CL	RP__CL	ਉਏ ਨੀ ਜਨਾਬ
9.3	Interjection	INJ	RP__INJ	ਬਹੁਤ
				ਬੜ੍ਹ
9.4	Intensifier	INTF	RP__INTF	ਨਹ ਨਾ ਵਗੈਰ
9.5	Negation	NEG	RP__NEG	ਥੋੜ੍ਹਾ ਬਹੁਤ ਕਾਫੀ ਕੁਝ ਇੱਕ
10	Quantifiers	QT	QT	ਥੋੜ੍ਹਾ ਕਾਫੀ ਕੁਝ
10.1	General	QTF	QT__QTF	ਇੱਕ ਦੇ
10.2	Cardinals	QTC	QT__QTC	ਪਿਹਲਾ ਦੂਜਾ
10.3	Ordinals	QTO	QT__QTO	
11	Residuals	RD	RD	
11.1	Foreign word	RDF	RD__RDF	$, &, *, (,)
11.2	Symbol	SYM	RD__SYM	., : ;
11.3	Punctuation	PUNC	RD__PUNC	
11.4	Unknown	UNK	RD__UNK	(ਪਾਣੀ-) ਧਾਣੀ (ਚਾਹ-) ਚੂਹ
11.5	Echowords	ECH	RD__ECH	ਆਇਆਂ ਕਿਰਆਂ

8. EXPERIMENTAL EVALUATION

The accuracy of any Part of Speech tagger is measured in terms of the accuracy i.e. the percentage of words which are accurately tagged by the tagger. This is defined as below:

Accuracy = Total no of words having correct tag / total no of words tagged

For evaluation of the proposed tagger, a corpus having texts from different online resources i.e. Punjabi websites were used. The outcome was manually evaluated from a linguistic expert to mark the correct and incorrect tag assignments. 2400 sentences having 10,000 words collected randomly.

Table 2: Results and analysis

corpus	No of words	No of unknown words	No of known words	No of correct tags assigned
Set1	5995	325	5670	5233
Set2	4007	344	3663	3369
Total	10002	669	9333	8602

The accuracy obtained was 92.16% (Ignoring the unknown words)

9. CONCLUSIONS AND FUTURE WORK

In this study, we proposed the implementation of N-Gram model for Part of speech tagging to one of the morphology rich language Punjabi. During experimental results we note that the general-Gram based method doesn't perform well due to

unknown words (foreign language words or due to spelling mistakes) problem. In future, we intend to develop novel methods to improve overall accuracy and specifically unknown words in Punjabi and other word-free languages. We aim to find out ways to improve the language model behavior without increasing the training corpus and by integrating linguistics knowledge.

10. 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] Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke, "Hindi Part of Speech Tagging and Chunking: A Maximum Entropy Approach", In Proceeding of the NLPAI Machine Learning Competition, 2006.

[6] 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

[7] Agarwal Himashu, Amni Anirudh," Part of Speech Tagging and Chunking with Conditional Random Fields" in the proceedings of NLPAI Contest, 2006

[8] Brants, TnT – A statistical part-of-speech tagger. In Proc. Of the 6th Applied NLP Conference, pp. 224-231, 2000

[9] Sanjeev Kumar Sharma and Dr G S Lehal "Improving Existing Punjabi POS tagger Using Hidden Markov Model"

[10] Jyoti Singh, Nisheeth Joshi and Iti Mathur in 2013 "Part Of Speech Tagging of Marathi text Using Trigram Model" in International Journal of Advanced Information Technology (IJAIT) Vol. 3, No.2, April2013 pp. 35-41.