Assigning the Correct Word Class to Punjabi Unknown Words using CRF

Sanjeev Kumar Sharma
Department of Computer Applications
DAV University, Jalandhar

ABSTRACT
Part of Speech tagging has a vital role in different fields of natural language processing. It can be defined as the process of assigning a tag or a label to a word according to its morphological or syntactical properties. The objective of this paper is to develop a POS tagger based on hybrid approach which is combination of rule based approach and CRF based approach. In this, the tagset used 36 tags which is proposed by TDIL for Indian languages.

Keywords
Natural Language Processing, Part of Speech Tagging, Rule based approach, CRF, Hybrid.

1. INTRODUCTION
Part of Speech tagging is one of the major task of Natural language Processing. It includes assigning grammatical labels of the word in the text. [25] Word class includes noun, pronoun, verb, adverb, adjective, etc. known as Part of Speech. Short labels are used for the sake of convenience, which are known as tags. For example, noun can be written as N and verb as V. It is generally the first step in the development of natural language processing applications. It takes a sentence as an input and assigns an appropriate label or tag to each word. These tagged words are used as input in various applications. It plays important role as processing becomes easier when the grammatical information about the word is known.

Very limited work has been done on Punjabi for Part of speech tagging. So, different approaches can be used for the development of efficient tagger. There are many approaches in natural language processing, which are mainly divided into three categories: Rule based, Statistical based and Hybrid approach. The Rule based approach uses linguistic rules to provide tags. This is the oldest approach in language processing. Statistical approach uses estimated probabilities to assign the most suitable tag to a word. This approach includes Hidden Markov Model (HMM), Maximum Entropy (ME), Support Vector Machine (SVM) and Conditional Random Field (CRF). The hybrid approach is the combination of the rule based and statistical approach. All these approaches are used in two methods i.e. supervised method and unsupervised method. The supervised tagging method is based on pre-tagged corpora. It includes the process of learning of the rules for tagging using annotated corpus. The unsupervised tagging method on the other hand do not require pre-tagged corpus.

Punjabi is an Indo-Aryan language spoken or understood by the people in India, Pakistan and other regions of the world by over 150 million people. Other members of Indo-Aryan family are Hindi, Bengali, Gujarati, and Marathi etc. Punjabi is written in 'Gurmukhi' script in eastern Punjab (India), and in 'Shahmukhi' script in western Punjab (Pakistan). Modern Punjabi vocabulary has been influenced by other languages, such as Persian, Sanskrit, Arabic, Urdu, Hindi and English.

2. PREVIOUS WORK DONE
Chirag Patel et al[14] proposed a CRF based model for Guajarati. The features which are provided to CRF are selected according to linguistic properties of Gujarati. The corpus is tagged manually because there is lack of resources. The tagset contains 26 different tags which is considered as standard for Indian languages. CRF learns from both tagged and untagged data. Due to the lack of flexibilities in features, new features are added in iteration to increase the accuracy. The accuracy obtained from this model is 89.90% which is improved to 92% after error analysis.

Himanshu Agrawal [1] proposed CRF model based POS tagger for Hindi. CRF is more suitable for large training data. The baseline performance of the system was obtained 78%. After error analysis, various features were added. Due to this, there is notable improvement. The best accuracy of the proposed system is 83%. The accuracy can be improves by increasing size of training data.

Kanak Mohnot [12] proposed hybrid approach based Hindi tagger. This tagger used 80000 words in the corpus and 7 different standard tags. The proposed system operated in two steps. Firstly the input data matches in the database. If they are present, then tags are provided. Secondly, if they aren't matched, then the HMM approach is used. After tokenization, if token isn't in the singular, then they are converted into a singular form. Various linguistic based rules are applied to find the appropriate tags. This tagger achieved the accuracy of 89.90%. It can be improved by further error analysis and by using more linguistic features.

Krishanpriya V et al.[18] developed a CRF based tagger for Malayalam. In this system, a standard tagged corpus of Linguistic Standard for Indian Language (IL) and BIS_Tagset are used. Basically, this system has been divided into three modules as Preprocessing, Training and Testing. This system achieved 85.7 % accuracy. The experiment was performed on both bigram and trigram. There is improvement observed with increase in grams.

Manchanda Blossom, Ravishanker [12] proposed an approach to find the POS tag of unknown words POS tagging in Punjabi using Trigram Model. Because of high information content, unknown words increase in number when words from different languages are used. All POS taggers suffer a significant decrease in accuracy because of unknown words. It is assumed that the unknown POS depends on the previous and next POS tags, and trigram probability is calculated to find maximum occurring combination. The POS tags for known words are taken from the tagged training corpus.

Singha Kh Raju, Purkayastha, Singha Kh Dhiran [17] proposed a model for POS tagging in Manipuri using HMM. As Manipuri has no tagged corpus, the system uses the small set of tagged sentence which is generated from Manipuri Rule-based Tagger. The system has the ability to assign tags to most of the lexical items in the test set. It gives the accuracy of 92% and it is clear that accuracy percentage was increased with increase in the size of the tagged corpus. The proposed system can be made more efficient by extending the bigram probability to trigram probability.

3. EXISTING POS TAGGER OF PUNJABI LANGUAGE

Gurpreet Singh Lehal et al. [8] proposed a rule based tagger for Punjabi language. This is based on the handwritten linguistic rules of the language. It includes an approximate 630 tags for various word classes, specific words and punctuations. Separate databases were designed to store the rules and to maintain the marked verbal operators. The proposed system achieved an accuracy of 80.29% including unknown words and 88.86% excluding unknown words.

Gurpreet Singh Lehal et al. [11] proposed Punjabi HMM based tagger. It uses bigram approach. The corpus used in this tagger was annotated by using existing tagger. It uses a Maximum Likelihood approach to determine the parameters, i.e. lexical probability and contextual probability. It provides the most suitable tag the word which maximizes the product of these probability. Viterbi algorithm is used to select the optimal tag from probabilities. This system obtained the accuracy of 90.11%.

Gurpreet Singh Lehal and Sanjeev K Sharma [9] proposed a maximum entropy based tagger for Punjabi language. This system is based on the Trigram model. It includes an approximate 630 tags for various word classes, specific words and punctuations. It also works to provide tags to Unknown words.

4. CONDITIONAL RANDOM FIELDS

Conditional Random Field is a statistical based approach which predicts sequences of labels or tags for the given input data [10]. CRFs are undirected graphical models, also known as random field, which is used to calculate the conditional probability p(x|y) of a possible output nodes y=(y1,…,y_n) given the input x=(x_1,…,x_2) which is also called the observation. A CRF in general can be expressed as

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(x)} \prod_{C \in c} \Psi_C(\vec{x}_C, \vec{y}_C)$$

It considers features such as neighboring words and word bigrams, prefixes and suffixes, capitalization, membership in domain-specific lexicons and semantic information from source.

CRF have been applied to a variety of domains, including text processing, computer vision, and bioinformatics. A problem in this model is computational expense of training.

5. FEATURES

The set of features that have been applied for POS tagging:

Table 1

Features	
Current Word	W_i
POS information	t_i
Word length	L_i
Previous word-next word	W_{i-2}, W_{i-1}, W_{i+1}, W_{i+2}
POS info of Previous word-next word	t_{i-2}, t_{i-1}, t_{i+1}, t_{i+2}

Context word feature: Preceding and following words of a particular word can be used as features.

Word suffix and prefix: Word suffix and prefix information is helpful to identify POS class.

Part of Speech (POS) Information: POS information of the current word with its previous word and next word are used as feature in the experiment.

Length of a word: Length of a word is used as feature of POS tagging.

Category information: The category i.e. root and word is used as features.

Gazetteer list: It contains words which can be name of person, place, words from other languages, etc. This is used when data do not exist in the training data.

6. SYSTEM ARCHITECTURE

There are mainly two modules, learning or training and Testing.

In learning phase, the system learns to predict tags from the input data. It requires a training file and template file to train the system. In this phase, annotated corpus is used. A model file is created by using the template file and training file. For this phase we use the following command

crf_learn template_file train_file model_file

Where template_file and train_file are prepared and model file is generated by crf_learn command.

In the testing phase, model file created in earlier phase is used. The testing file is also required to be in the standard column format.

For this the following command is used

crf_test –m model_file test_file

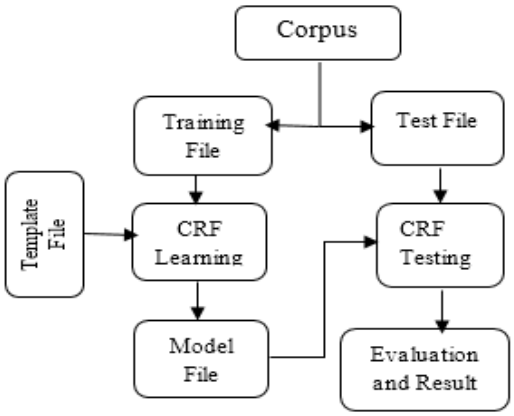

Fig 1 System Architecture

Where model_file is generated in learning phase and test_file contains the data to be tested.

7. EXPERIMENT AND RESULTS

The corpus used in this system contains approximately 38k to 42k words. From this 2/3 of the sentences are used for training and the remaining are used in testing of the proposed system. We have used CRF++-0.58 toolkit in the proposed system. The tagset used in this system contains 36 tags. This is proposed by TDIL (Technical Development of Indian Languages) for Indian languages.

To calculate performance of the system precision and recall are calculated. These are calculated as

$$Precision\ (P) = \frac{No.of\ correct\ POS\ tags\ assigned\ by\ the\ system}{No.of\ POS\ tags\ assigned\ by\ the\ system}$$

$$Recall\ (R) = \frac{No.of\ correct\ POS\ tags\ assigned\ by\ the\ system}{No.of\ POS\ tags\ in\ the\ text}$$

Table 2: Experimental Results

Corpus	**Total number of words**	**No of unknown words (not tagged by the system)**	**No of known words**	**Existing HMM based system**	**Proposed system**
				No of correctly disambiguated tags	**No of correctly disambiguated tags**
Articles	6594	357	6237	5756	6171
News	3205	275	2930	2695	2876
Stories	8461	62	8399	7741	8367
Novel	3762	316	3446	3282	3400
EBook	2347	25	2322	2214	2298

Table 3: Experimental Results

Corpus type	**Existing HMM based system**					**Proposed Hybrid System**				
	A	**B**	**C**	**Precision**	**Recall**	**A**	**B**	**C**	**Precision**	**Recall**
Articles	5756	412	0	100%	92.8%	6171	0	66	98.9%	100%
News	2695	200	0	100%	92.5%	2876	0	54	98.1%	100%
Stories	7741	558	0	100%	92.7%	8367	0	32	99.6%	100%
Novel	3282	148	0	100%	95.4%	3400	0	45	98.6%	100%
EBook	2214	99	0	100%	95.5%	2298	0	24	98.9%	100%

8. CONCLUSION

In this work we have proposed a CRF approach based part of speech tagger for Punjabi language. CRF is a statistical approach used by different authors for different languages but has never been used in Punjabi language. There is significant improvement in the accuracy, as the previous proposed HMM based tagger achieved accuracy was 90.11%.

9. REFERENCES

[1] Agrawal Himanshu, Mani Anirudh, "Part of Speech Tagging and Chunking with Conditional Random Fields", NLPAI Machine Learning Contest 2006.

[2] Aniket Dalal, Kumar Nagaraj et al. "Hindi part-of-speech tagging and chunking: A maximum entropy

approach." Proceeding of the NLPAI Machine Learning Competition (2006).

[3] Francis Merin, Nair K N Ramachandran, "Hybrid Part of Speech Tagger for Malayalam", 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 978- 1-4799-3080-71 14/$3l.00 ©20 14 IEEE.

[4] Garg Navneet, Goyal Vishal, Suman Preet," Rule Based Hindi Part of Speech Tagger", Proceedings of COLING 2012: Demonstration Papers, pages 163–174, COLING 2012, Mumbai, December 2012.

[5] Joshi Nisheeth, Darbari Hemant, Mathur Iti, "HMM based POS tagger for Hindi", Jan Zizka (Eds): CCSIT, SIPP, AISC, PDCTA – 2013, © CS & IT-CSCP 2013.

[6] Klinger Roman, Tomanek Katrin, "Classical Probabilistic Models and Conditional Random Fields", Algorithm Engineering Report, TR07-2-013, December 2007, ISSN 1864-4503.

[7] Kumar Dinesh, Josan Gurpreet Singh, "Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 6– No.5, September 2010.

[8] Lehal Gurpreet Singh, Sharma Sanjeev K, "Maximum Entropy Method for POS Guessing Of Punjabi Unknown Words", IMS - International Conference on Information and Mathematical Sciences 2013.

[9] Lehal Gurpreet Singh," A Survey of the State of the Art in Punjabi Language Processing", Language in India, www. languageinindia.com, Strength for Today and Bright Hope for Tomorrow", Volume 9: 10 October 2009, ISSN 1930-2940.

[10] Lehal Gurpreet Singh, Sharma Sanjeev K, "Using Hidden Markov Model to Improve the Accuracy of Punjabi POS Tagger", 978-1-4244-8728-8/11/$26.00 ©2011IEEE.

[11] Manchanda Blossom, Ravishanker," To find the POS tag of unknown words in Punjabi language", An International Journal of Engineering Sciences ISSN: 2229-6913 Issue July 2011, Vol. 1.

[12] Mohnot Kanak, Bansal Neha, Singh Shashi Pal, Kumar Ajai, "Hybrid approach for Part of Speech Tagger for Hindi language", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 4, Issue 1, February 2014.

[13] Nadkarni Prakash M, Machado Lucila Ohno, Chapman Wendy W," Natural language processing: An introduction", J Am Med Inform Assoc 2011; 18:544e551. DOI: 10.1136/amiajnl-2011-000464, Published by group.bmj.com on October 5, 2011.

[14] Patel Chirag, Gali Karthik, "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 117–122, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing.

[15] Singh Thoudam Doren, Ekbal Asif, Bandyopadhyay Sivaji, "Manipuri POS Tagging using CRF and SVM: A Language Independent Approach", 6th International Conference on Natural Language Processing, 2008.

[16] Singha Kh Raju, Purkayastha Bipul Syam, Singha Kh Dhiren, "Part of Speech Tagging in Manipuri: A Rule-based Approach", International Journal of Computer Applications (0975 – 8887), Volume 51– No.14, August 2012.

[17] V Krishnapriya, P Sreesha, T.R. Harithalakshmi, T.C. Archana, Vettath Jayasree N, "Design of a POS Tagger using Conditional Random Fields for Malayalam", First International Conference on Computational Systems and Communications (ICCSC), Trivandrum, 978-1-4799-6013-2/14/$31.00 ©2014 IEEE.

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