# Word Class Prediction of Ambiguous and Unknown Words of Punjabi Language Using Bi-gram Methods

SanyamSood
Assistant professor
CSE
NWIET Moga

Vishal Arora
Assistant professor
CSE
SBSSTC Ferozepur

Sanjeev Kumar Sharma
Assistant Professor
Department of Computer Science and Application
DAV University Jalandhar

***Abstract: - Ambiguous and unknown words are found in every language. Ambiguous words are the words having different meaning in different sentences depending upon the context of the sentence. Assigning the correct word class to these ambiguous words is the fundamental task in almost all the NLP applications. A lot of work has been done on this and a lot of work is still to be done. Many statistical and rule based techniques has been applied to assign the correct word class to the word having ambiguous word class. Most commonly used statistical techniques are HMM (Hidden Markov Model), SVM (Support Vector Machine), ME (Maximum Entropy), CRF (Conditional Random Field) and N-gram based techniques. In this research paper a bigram technique has been discussed to assign the correct word class to the ambiguous and unknown words of Punjabi language. A tag set proposed by TDIL has been used to assign the correct word class to the ambiguous and unknown words.***

***Keywords-***Ambiguous words, word class, Unknown words, Bi-gram technique, TDIL proposed Punjabi tag set.

**INTRODUCTION**

Ambiguous words and unknown words are the issues which need to be solved in almost all the languages. Ambiguous words are those words which have different meaning when used in different sentences depending upon the context of the sentences. Unknown words are the words which are not present in the morph of the language. These words are either misspelled words of Punjabi language or words of foreign language. Both these are major hurdle in further processing of the language and hence need to be rectify before further processing of the language. Here the word class means the grammatical information of the word. The word class is assigned in the form of tag called POS (part of speech) tag. So a process is needed to assign the appropriate word class to the ambiguous and unknown word. This process is also called POS tagging. This process is necessary for further processing of the sentence like sentence identification, grammar checking,and machinetranslation and in NER etc. and hence it is the fundamental task performed in almost all the NLP applications.

## Introduction to Punjabi Language:

Punjabi language is one of the most spoken languages in India and belongs to the Indo-Aryan family of languages. Indo-Aryan languages are also called Indic languages. Hindi, Gujarati, Bengali and Marathi etc. are some of the other members of this family. Most of the Punjabi speakers belongs India, Pakistan, USA, Canada, England etc. Punjabi is known to be the 13th most spoken language in India and is the official language of the state of Punjab in India. Punjabi is written in two scripts, one is "Gurmukhi" script and other is "Shahmukhi" script.

## Word classes in Punjabi language:

In Punjabi language there are 11 word classes i.e. noun, pronoun, and adjective, adverb, verb, cardinal, ordinal, conjunction, preposition, particle and verb part . All the Punjabi word lies in one or more than one of these 11 word classes. The purpose of POS tagging is to assign the correct word class to each word in the sentence.

## Ambiguous and unknown words in Punjabi language:

As discussed in introduction section the ambiguous words can have different word classes in different sentences or in different context. This can be explained by the following example:

ਮਾਸਟਰਦਾਸਾਈਕਲਗੁੰਮਹੋਗਿਆਹੈਤੇਉਹਇਸਨੂੰਲੱਭਰਿਹਾਹੈ।

(māsṭ ardāsāīkalgummhōgiāhaitē uh is nūṃlabbhrihāhai .)

The morph will assign the word class to each word. After assigning the word class the output is:

( ਮਾਸਟਰ_NN ਦਾ_PPI ਸਾਈਕਲ_NN ਗੁੰਮ_VBMAXਹੋ_VBMAX ਗਿਆ_VBOP ਹੈ_VBAXBST1 ਤੇ_CJC ਉਹ_PND|IJ ਇਸ_PND ਨੂੰ_PPU ਲੱਭ_VBMAX ਰਿਹਾ_VBOR ਹੈ_VBAX ।_Sentence )

In above example the word ਉਹ(uh) has been assigned ambiguous word class by the morph. As shows above, wordਉਹ(uh) can be used a demonstrative pronoun (PND) or as an interjection (IJ) in the sentence. The main task of our research is to assign such a system that could assign the appropriate tag to each word out of allpossible tags assigned by morph.

The unknown words are those which were not present in the morph and hence have been assign an unknown tag by the system. E.g

ਚੰਗੀ ਬੋਲਚਾਲ ਨਾਲ ਅਸੀਂ ਹਰ ਕਿਸੇ ਨੂੰ ਆਪਣਾ ਮਿੱਤਰ ਬਣਾ ਲੈਂਦੇ ਹਨ।

(caṅ gībōlcālnālasīṃharkisēnūṃāpaṇ āmittarbaṇ ālaindēhan.)

( ਚੰਗੀ_AJI ਬੋਲਚਾਲ_Unknown ਨਾਲ_AVU ਅਸੀਂ_PNP ਹਰ_AJU ਕਿਸੇ_PNI ਨੂੰ_PPU ਆਪਣਾ_AJIMSD ਮਿੱਤਰ_NN ਬਣਾ_VBMAX ਲੈਂਦੇ_VBOP ਹਨ_VBAX ।_Sentence )

In above sentence, ਬੋਲਚਾਲ(bōlcāl) is the unknown word as “Unknown” tag has been assigned to it. This is again a problem for processing the natural language.

## Related Work:

There are basically three techniques used for part of speech tagging. 1) Rule based method 2) Statistical based method and Neural network based method. Besides these three a hybrid method is also used. This hybrid method is the combination of two or three of above mention techniques. In rule based technique different hand written rules are used to remove ambiguity. These rules are developed manually. Therefore thorough knowledge of language is required to develop the rules. This rule based technique has been used by Sreeganesh (2006) for Telugu language; another rule based POS tagger was developed for Punjabi language by Mandeep Singh Gill, Gurpreet Singh Lehal (2008). Statistical method is another technique commonly used for part of speech tagging. Most commonly used statistical methods are support vector machine (SVM) used by Ekbal and S. Bandyopadhyay (2008) for Bengali language; V.Dhanalakshmi et al. (2008)for Tamil language, M Anandkumar, Vijaya M.S, Loganathan R, Soman K.P, Rjendran S (2008) ;SindhiyaBinulal et al. for POS tagging of Telugu language. Antony P.J et al. for Malyalam language. Hidden markov model based technique used by Manish Shrivastava&Pushpak Bhattacharyya for POS tagger for Hindi language; Manju K et al. for Malayalam language; NavanathSaharia et.al for Assamese; Sanjeevkumar Sharma et al. (2011) for Punjabi Language; Ekbal, S. Mondal et al. for Bengali language. Maximum entropy based technique was used by AniketDalal et al. for Hindi language; Ekbal et al. (2008) for Bengali language. Conditional Random Field based technique has been used by Ravindran et al. and Himanshu et al. for POS tagging and chunking of Hindi language; other Indian languages on which this CRF technique has been applied are Bengali and Manipuri. Neural network based technique has been used by Ankur Parikh for Hindi Language. In hybrid based approach used a combination of rule based and HMM based technique has been used by Arulmozhi P et al. for development of Tamil POS tagger; Chirag Patel and KarthikGali [8] used a combination of rule based method and CRF for Gujarati POS tagger.

## Existing system for Punjabi Language:

Currently there are two systems exist for word class disambiguation in Punjabi language. These system have been developed by using two different techniques i.e. rule based and statistical based. Rule based system was developed by Mandeep Singh Gill et al. (2008) [17] as a sub part of grammar checker. A large tag set of more than 630 tags was used and exhaustive set of rule developed by a linguistic person were used. Statistical based system was developed by Sanjeevkumar Sharma et al. (2011). The Hidden Markov Model (HMM) based approach was used to disambiguate the tags. Viterbi algorithm was used to implement the Hidden Markov Model.

## Tag set proposed by TDIL:

Depending on some general principle of tag set design strategy, a number of POS tag sets have been developed by different organizations. For POS annotation of texts in Punjabi, we have used tag set proposed by TDIL (Technical Development of Indian Languages). There were 36 tags proposed by TDIL for Punjabi language.

## Introduction to Bi-Gram

Bi-gram is a probability based technique in which the correct tag to word having more than one tag is assigned on the basis of its probability with the previous tag. These probability values are pre calculated from a training corpus. The accuracy of this technique increases with increase in training corpus.

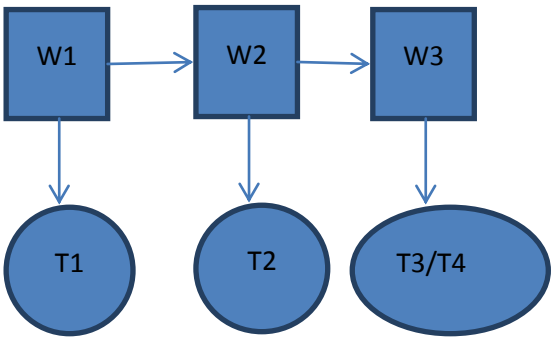


**Figure:-1 bigram model**

Consider the above diagram. In this diagram first two words have single word class and the third word has an ambiguity i.e. has two word classes. In bigram approach all the combinations of ambiguous word classes with previous word class are created i.e. T2_T3 and T2_T4. Then each word class pair is assigned a probability from bigram probability file. From these two word class pairs, the one having maximum probability is selected and the word classes from this pair are assigned to respective words.

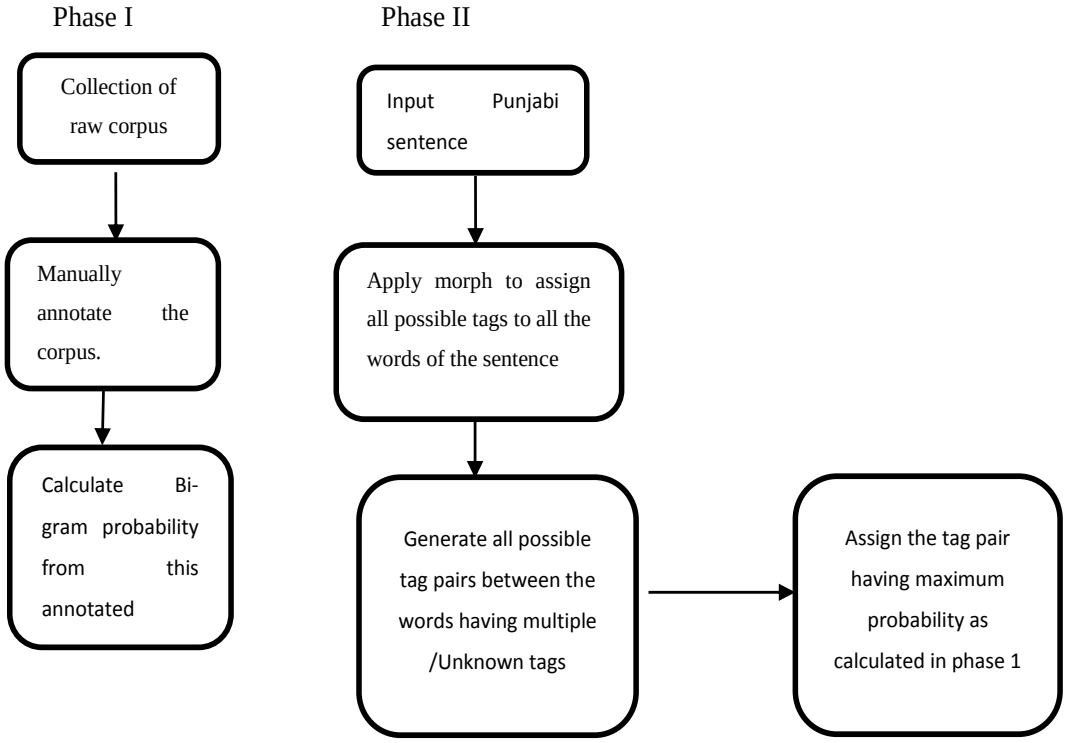


Figure 2: Flow Chart of Word Class disambiguation Using Bi-gram Approach

**METHODOLOGY**

**Step 1**:- collection of raw corpus.

A large accurate corpus of nearly 50 pages containing nearly 2000 sentences and approximately 16,000 words has been collected from the following various Punjabi newspaper websites and other reliable online resources.

**Step 2**:- Annotation of corpus.

The corpus was manually annotated with the help of linguistic person. For annotation the word classes in the form of tags proposed by TDIL were used.

**Step 3**:- Calculation of bi-gram probability

The bi-gram probability has been calculated by using the following formula:

$P(t_i\_t_j) = C(t_i\_t_j) / C(t_i)$
In above formula $P(t_i\_t_j)$ is the probability of occurrence of tag tj after tag ti. $C(t_i\_t_j)$ is the total number of times the tag pair $t_i_t_j$occurs in the training corpus and $C(t_i)$ is the number of times the tag $t_j$occurs in the training corpus. A sample bi-gram probability file has been shown in table1.

**.Table 1: Sample Bi-gram Probability File**

| Tag1/Tag2 pair | Probability |
|---|---|
| N_NN/V | 0.190476 |
| V/V_VM | 0.005376 |
| V_VM/RP | 0.040698 |
| RP/PR_PRP | 0.016484 |
| PR_PRP/N_NN | 0.01066 |
| N_NN/PSP | 0.470769 |
| PSP/N_NN | 0.028161 |
| N_NN/V_VM | 0.135289 |

## Experimental Evaluation

The accuracy of our system has been calculated in the following way:

% age of word class disambiguation=$\frac{number\ of\ correcly\ disambiguatedwords}{total\ number\ of\ ambigous\ words}$

Similarly accuracy of unknown word prediction can be calculated as:

% age of Unknown words prediction=$\frac{\text{number of Unknown word correcly identified}}{\text{total number of Unknown words}}$

For evaluation of the proposed POS tagger, a corpus having texts from different online resources i.e. Punjabi websites were used. The outcome was manually evaluated through a linguistic expert to mark the correct and incorrect disambiguate tags. The results obtained have been given in Table 2.

**Table 2: Experimental Result**

| Corpus | Total number of words (A) | Number of words having ambiguous word class (B) | Number of words correctly disambiguated | %age accuracy of word class disambiguation system (B/A)*100 | Number of unknown words (assigned as "Unknown" by the morph) ( C ) | Number of Unknown words correctly assigned word class (D) | %age accuracy of Unknown word detection system (D/C)*100 |
|---|---|---|---|---|---|---|---|
| Corpus1 | 2003 | 1611 | 1523 | 94.53 | 110 | 102 | 92.72 |
| Corpus2 | 2001 | 1567 | 1489 | 95.02 | 45 | 40 | 88.8 |
| Corpus3 | 2009 | 1754 | 1620 | 92.36 | 98 | 92 | 93.87 |

## Conclusion:

Efforts to improve the accuracy of Punjabi word class disambiguation system have been done by using bigram technique along with unknown word tagging component. The tagset proposed by TDIL has been used.It has been observed that there is significant improvement in the accuracy of word class disambiguation. Our proposed tagger equipped with unknown tag guesser component shows an accuracy of 92-94% whereas the existing HMM based POS tagger was reported to give an accuracy of 85-87% [1]. This significant improvement is due to reduction in the tagset from more than 630 tags to 36 tags and introduction of unknown tag guesser component.The unknown word tag guesser component also gave an accuracy of 88-94%.

## Future scope:

Further extension in the work can be done by using a hybrid model i.e. combination of more than one models. The same approach can also be implemented for other similar Indian languages like Hindi, Bengali, and Telugu etc. further improvement can be done by applying different approached to guess the word class of unknown words.

**REFERENCES**

[1] Sharma S.K, Lehal G.S (2011) "Using HMM to Improve accuracy of Punjabi POS tagger"2011 IEEE International Conference on computer science and Automation Engineering. Shanghai (China)
[2] 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.
[3] AniketDalal, 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.
[4] Ankur Parikh (2009), "Part-Of-Speech Tagging using neural network", Proceedings of ICON-2009: 7th International Conference on Natural Language Processing.
[5] 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.
[6] AnupamBasu, Ray, *RanjanPradipta*, 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).
[7] 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.
[8] Avinesh PVS and GaliKarthik (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.
[9] Chirag Patel and GaliKarthik (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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] Ganesan M (2007), "Morph and POS Tagger for Tamil" (Software) Annamalai University, Annamalai Nagar.
[16] G.SindhiyaBinulal, 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.
[17] 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.
[18] Mandeep Singh Gill, Lehal G.S. (2008) "Grammer Checking System for Punjabi" Coling 2008: companion volume Posters and Demonstrations pages 149–152 Manchester.
[19] 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.
[20] Manjuk,SSoumya , 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 .
[21] NavanathSaharia, Das Dhrubajyoti, Sharma Utpal, KalitaJugal (2009), "Part of Speech Tagger for Assamese Text", Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore, pp. 33–36.
[22] http://tdil.mit.gov.in/