GJCRT : Global Journal of Computer Science Research & Technology
ISSN (online): 2349- 0608

Development Of POS Tagger For Punjabi Language

Kamaljeet K Batra
Assistant Professor
DAV College, Amritsar
kamaljit_batra@gmail.com

Abstract

A part-of-speech tagger is a system that uses context to assign parts of speech to words.POS tagging involves process for disambiguating the part-of-speech information for such ambiguous words by taking into consideration the context information. The paper describes a statistical approach in building a Part-of-Speech(POS) tagger for Punjabi language. The steps involved are tokenization, morph analyzing, assigning tags and disambiguating tags. The accuracy of the tagger comes out to be 91.5%.

Keywords:

Tagger, Morph Analyzer, Tokenization

I. INTRODUCTION

Part-of-Speech (POS) tagging is the task of assigning a POS tag to every word in the input text. Tag is defined as a part-of-speech label. As natural languages are ambiguous in nature, i.e. have more than one POS tag, therefore, POS tagging also involves process for disambiguating the part-of-speech information for such ambiguous words by taking into consideration the context information. A part-of-speech tagger is a system that uses context to assign parts of speech to words. In a typical NLP application, the input for POS tagging is usually the outcome of either morphological analysis or simple lexicon look up, and that output is often ambiguous, i.e. having all the possible POS tags for each word. Therefore, the job of a typical POS tagger is to select the most appropriate POS tag for a given word from the list of possible POS tags for that word. It decides this based on the surrounding words and/or their POS tags. The basic parts of a POS tagger are POS tagset (i.e. tagging scheme) and disambiguation approach. The selection of features to be incorporated in the POS tagset depends upon the application for which POS tagging is to be used. The common approaches followed for disambiguation can be broadly categorized as rule-based and statistics-based. A combination of these two approaches was also attempted and is reported to be more successful. Given below are some of the existing POS tagging systems, with focus on the disambiguation approach followed.

This paper describes a statistical approach in building a Part-of-Speech(POS) tagger for Punjabi language.

II. REVIEW OF LITERATURE

Church (1988) presented a part-of-speech tagging program developed at Bell Labs. This program follows a stochastic approach and performs tagging by maximizing the lexical and contextual probability for a given sentence. Lexical probability is calculated as frequency of occurrence of a given word as a particular part-of-speech divided

by frequency of occurrence of that word, in the training corpus. This part-of-speech tagger is reported to achieve accuracy between 95-99%.[2]

Cutting et al. (1992) described a HMM based part-of-speech tagger developed at Xerox research, known as **Xerox tagger**. Authors claimed that by using only a lexicon and small amount of unlabelled text, the system achieved accuracy more than 96%. [3]

Brill (1992) applied a rule-based approach to automatic part-of-speech tagging. It automatically acquires rules and tags from the given training data. This approach requires a small set of meaningful rules as opposed to the large tables of statistics required for stochastic taggers. The rules can be easily understood, and therefore updation is fast and easy. [12]

Marcus et al. (1993) in their research paper presented a part-of-speech tagging process involved in building up a 4.5 million word corpus of American English, as part of **Penn Treebank** project at University of Pennsylvania. Skeletal syntactic structure is also being marked in this corpus. The tagging process used for building up this corpus was two stage – in the first stage an automatic process of tagging the corpus was used and in second stage the errors in the automatic tag assignment were corrected manually.[5]

Schmid (1994) applied neural network based learning to part-of-speech tagging of English. The author noted that this approach produces similar accuracy results as for trigram based tagger and it is more accurate than HMM based tagger. Schmid (1994b) applied decision trees to learn part-of-speech tags. It is noted that the tagger (known as **TreeTagger**) produces 96.36% accuracy level when tested on Penn Treebank corpus which is better than 96.06% produced by a trigram tagger on the same data.[6]

Chanod and Tapanainen (1995) applied both statistical and rule-based approaches to part-of-speech tagging of French. For statistical approach, Xerox tagger developed by Cutting et al. (1992) is used. The accuracy reported for this tagger, when applied to French, is 96.8%.[7]

III. STEPS FOLLOWED FOR BUILDING POS TAGGER

A Tokenization

Tokenization refers to the process of breaking the text into tokens or words using punctuation marks and spaces as delimiters. A simple program could separate the text into word and punctuation tokens simply by breaking it up at white-space and punctuation marks. White-space is a fairly reliable indicator of a Punjabi token boundary.

Eg A given sentence

minstr nUM mwr dyx dI DmkI idSqI geI can be divided into following tokens

minstr, nUM, mwr, dyx, dI , DmkI, idSqI, geI

B Morph Analyzing

Morphological Analysis provides information about a word's semantics and the syntactic role it plays in a sentence. For each word form of a text, the analysis system determines its root, part of speech, and – if appropriate - its gender, case, number, person, tense, and comparative degree. Indian languages are characterised by a very rich system of inflections, derivation and compound word formation for which a standard morphological analyser is

necessary to deal with any type of text. The number of words are derived from a given root word by some specific syntactic rules. Morphological analysis is the identification of a stem-form from a full word- form. For example, the analyzer must be able to interpret the root form of "**muMfy"** as **"muMfw**" and its GNP(Gender-Number-Person) information A Punjabi morph analyzer has been developed using the morph database developed at "Advanced centre for technical development of Punjabi language, Literature and Culture" which analyzes the exact grammatical structure of the word. The morph database used in the system includes, the information about every word in Punjabi, with the information about its gender, number, person, case, tense etc. Every inflected word also contains the root word from where it is derived. The database contains more than one lac words from which 63,000 are the inflected nouns which are derived from about 18,000 root nouns. The database contains the grammatical category of each word and also the inflected words it can form. From this database, the information is retrieved and each word of the sentence is tagged with a specific tag sets. In Punjabi grammar, the parts of speech include noun, verb , adjective, adverb, pronoun, preposition etc. [13,14]

Table 1 shows the main Table, 2 shows the design of morph database for nouns.

Table 1: Morph Database Design for Main table

Field Name	**Description**
Word	Inflected word
Category	Grammatical Category

Table 2: Morph Database Design for nouns

Field Name	**Description**
Word	Inflected word
Root	Root word
Gender	Gender information
Number	Number information
Case	Case information

Sample Database Entries for nouns

Word	**Root**	**Gender**	**Number**	**Case**
aumIdvwr	**aumIdvwr**	Masculine	Plural	Direct
aumIdvwr	aumIdvwr	Masculine	Singular	Oblique
aumIdvwr	**aumIdvwr**	Masculine	Singular	Direct
aumMgwN	**aumMg**	Feminine	Plural	Oblique
Aumr	**Aumr**	Feminine	Plural	Direct
Aumr	**Aumr**	Feminine	Singular	Oblique

Word	Root	Gender	Number	Case
bwlVI	bwlVI	Feminine	Singular	Direct
bwlVI	bwlVI	Feminine	Singular	Oblique
bwlVIEwN	bwlVI	Feminine	Plural	Oblique

C Assigning Tags

The classification in the morph analyzer is used to develop an appropriate tagging scheme. The hierarchy of parts of speech allows the development of meaningful tags easily expandable to include more details and precision about the Punjabi units whenever needed. Each tag contains the information about the grammatical category of word, gender, number, person and the case. For each type of tag, separate information is provided for separate grammatical categories. Tags prepared after analyzing the sentence is arranged in the form 'grammatical category-gender-person-number-case-tense-phrase-type'. The fields not applicable to a particular category are left blank.

The parts of tag separated by hyphen('-') are arranged at following positions

1 - 2 - 3 - 4 - 5 - 6 - 7 - 8

grammaticalcategory-gender-person-number-case-tense-phrase-type'.

Table 3: POS Tagset for Punjabi

Position	Category/Feature	Symbol-Description
1	**Word class**	**n-Noun**
2	Gender	m-Masculine, f-Feminine, b-Both
4	Number	s-Singular, p-Plural, x-Both
5	Case	d-Direct, o-Oblique, v-Vocative, a-Ablative, l-Locative, i-Instrumental
1	**Word class**	**pp-Personal Pronoun**
2	Gender	m-Masculine, f-Feminine, b-Both
3	Person	f-First, s-Second, t-Third
4	Number	s-Singular, p-Plural
5	Case	d-Direct, o-Oblique, v-Vocative, a-Ablative, d-Dative

In the above table, first column shows the position of grammatical categories and features related to the category and the last column shows the description of symbols related to the features. The total tags came out to be 658.

Example Tags

Tags for the word **'Brw'** (Bhra) **are 'n-m- -s-d- - - -', 'n-m- -p-d- - - -','v-x-s-s- -f-x- -'.** The above tag for the word shows that it can be used as noun with masculine gender, singular as well as plural and in direct case. It can also be used as verb with any gender, singular, second person, and future tense. In Punjabi language, a word can have

number of tags as a particular word can be used in number of ways. It first checks the category of each word from the database and then adds Gender, Number, Person or Tense information to it.

For example in case of Nouns

Tagger gives information of Gender, Number and Case (n-G- -N-C- - - -)

Tag for the word 'Brw' (Bhra) – n-m- -s-d- - - - (noun-masculine-singular and direct case), In case of nouns person information is not in use.

Similarly for personal pronouns

'myrI' (meri)– pp-f-f-s-d- - - - (personal pronoun – feminine, first person, singular, and in direct case)

D Ambiguity Resolution for tags

Ambiguity exists when a particular word can have number of tags of different grammatical category or with different GNP information. There are three general approaches to deal with the tagging problem:

Rule-based approach: consists of developing a knowledge base of rules written by linguists to define precisely how and where to assign the various POS tags. Several grammatical rules gives some signs to distinguish between type of word, and others signs are deduced from other features (number, gender, preposition, and conjunction...etc.) During tagging process, the context and word form features are looked up for each word in the text. Information about surrounding words is used, two words of the right context and two words of the left context [8]. The rules considering the tags for surrounding words are used for resolving ambiguities at different levels. Before the step of ambiguity resolution, each word is attached with number of tags. Since a particular word may have number of tags, there is need to check which tag is applicable to a particular word in a sentence. For this purpose, there is need to apply certain rules depending upon the grammatical category, number, gender or other information.

Statistical approach: It consists of building a trainable model and to use previously-tagged corpus to estimate its parameters. Once this is done, the model can be used to automatically tagging other texts. Successful statistical taggers were built during the last years and are mainly based on Hidden Markov Models (HMMs). Generally, the most probable tag sequence is assigned to each sentence following the Viterbi algorithm. Part of Speech (POS) tagging is to find the sequence of POS tags T = {t1, t2, t3, . . . tn} that is optimal for a word sequence W = {w1, w2, w3 . . .wn}. The tagging problem becomes equivalent to searching for argmaxT P(T) * P(W | T), by the application of Bayes' law. The probability of the tag i.e., P(T) can be calculated by Markov assumption which states that the probability of a tag is dependent only on a small, fixed umber of previous tags. For tri-gram model, i.e., the probability of a tag depends on two previous tags, and then we have, P(T) = P(t1) * P(t2 | t1) * P(t3 | t1, t2) * P(t4 | t2, t3) * . . . * P(tn | tn - 2 , tn - 1). The Viterbi algorithm (Viterbi, 1967) allows us to find the best T in the linear time. The idea behind the algorithm is that of all the state sequences, only the most probable of these sequences need to be considered. The trigram model has been used in the present work. [9,10]

Hybrid approach: It combines the rule-based approach with a statistical one. Most of the recent study uses this approach as it gives better results.

IV OUR APPROACH

This tagger uses a hybrid approach. In this approach, a form of combination between statistical and linguistic approaches will be employed, so that the processing is performed in two levels.

A Ambiguity Resolution for words with different grammatical categories

For resolving ambiguity for different grammatical categories, an appropriate statistical model based on the internal structure of the Punjabi sentence is used to recognize the morphological characteristics of the words for the entered text. The use of the linguistic internal structure of the Punjabi sentence will allow us to identify logical sequences of words and consequently their corresponding tags. Since the probability of certain word (or its tag) occurrence depends on the words preceding it in a given context. The HMM is the best suitable statistical model to keep track of this history. Each state of the HMM is represented by a possible tag in the lexicon and the transitions between states (tags) are governed by the syntax of the sentence. Transition probabilities are calculated using a smoothed tri-gram.

E.g..

P: auhdI bolI bhuq ruSKI sI

T: auhdi boli bahut rukhi see

In the above sentence, the word '**bolI**' (boli) has two tags, one show that it is noun and the other shows that it is verb. In the first sentence, the word `bolI` (boli) is preceded by demonstrative pronoun and probability of its occurrence as noun is more than verb, so the tag for noun is attached with this word.

But, if the sentence is

P: auh swfy Gr Ew ky bhuq auScI bolI

T: auh sadey ghar aa ke bahut uchi boli

Here the word '**bolI**' (boli) is used as verb, as it is preceded by two adverbs

Algorithm

Step 1: For each token in the given sentence perform step2 to step5

Step 2: For each tag of i^{th} token perform step 3

Step 3: If it is start word, calculate the probability for all the existing tags by considering the tag and word at start.

$$P(W_i|T_i)=freq(t_i,w_i)/freq(t_i)$$

Else if i^{th}word is present at 2^{nd} position in the sentence

$$P(W_i|T_i)=freq(t_{i-1},t_i,w_i)/freq(t_{i-1},t_i)$$

else

$$P(W_i|T_i)=freq(t_{i-2},t_{i-1},w_i)/freq(t_{i-2},t_{i-1},t_i)$$

End if

Step 4: Find the tag with maximum probability from the probabilities of all tags

Step 5: Set that tag to be the i^{th} tag

Sample entries of words with more than one tag

augrvwdI (inflected adjective or noun)

1. augrvwdI ny mYnUM mwr dyx dI DmkI idSqI (noun)

2. auh augrvwdI hY (adjective)
3. augrvwdI ny myry pqI nUM cwkU ivKwieEw (noun)
4. augrvwdI ivruSD kys drj kIqw igEw hY (noun)
5. augrvwdI kuVI nUM cuSk ky lY gey (noun)

The probability for this word to be noun if at start is 0.007 and adjective at start comes out to be 0. So the probability for occurrence of this word is noun if at start. Similary probabaility is calculated if preceded by demonstrative pronoun and in 2 sentence, tag for the word comes out to be adjective.

bdlw (noun or verb)

1. mohn ny borI bdlw leI (verb)
2. mYN auhnwN qoN Ewpxw bdlw jLrUr lYxw hY (noun)

The probability of occurrence of the word as verb preceded by a noun and postposition in sentence 1 is more as compared to noun and if preceded by reflexive pronoun and postposition, its occurrence is noun.

B Ambiguity Resolution for morphological information

Second level of ambiguity exists when the word has number of tags with same grammatical category, but with different Gender, Number, Person or Case. For resolving such type of ambiguity, rule based approach is used. Rules have been built for this purpose.

Some examples with the rules

1. A word has number of tags that shows a particular word as noun, but can be used as singular or plural as tag for the word '**bMdy**' (bandey) is, '**n-m- -s-o - - -', 'n-m- -p-d- - - -'**

The tagged word can be noun in singular or a noun in plural.

E.g.. In the sentence,

P: 'bhuq swry **bMdyy lVn Ew gey'**

T: bahut sarey bandey ladhan aa gaye

In this case we should select the tag **'n-m- -p-d- - - - -',** and its appropriate word in English is '**men**', whereas in the case

P: 'auh bMdy ny kuVI nUM CyiVEw',

T: auh bandey ne kudhi nu chereaa

the tag for 'bMdy' (bandey) should be '**n-m- -s-o- - - -'** and its appropriate meaning is '**man**'. Such type of ambiguity can be resolved by considering the number i.e. Singular or plural of the auxiliary verb or the main verb present in the sentence. In the first sentence, 'gey' (gaye) is specified as auxiliary verb with plural attribute, whereas in second sentence 'CyiVEw' (chereaa) is specified as verb with singular attribute.

2.Similarly there are numbered tags for demonstrative pronouns.

P: ieh myrI ikqwb hY

T: ieh meri kitab hai

E: this is my book

P: ieh myrIEwN ikqwbwN hn

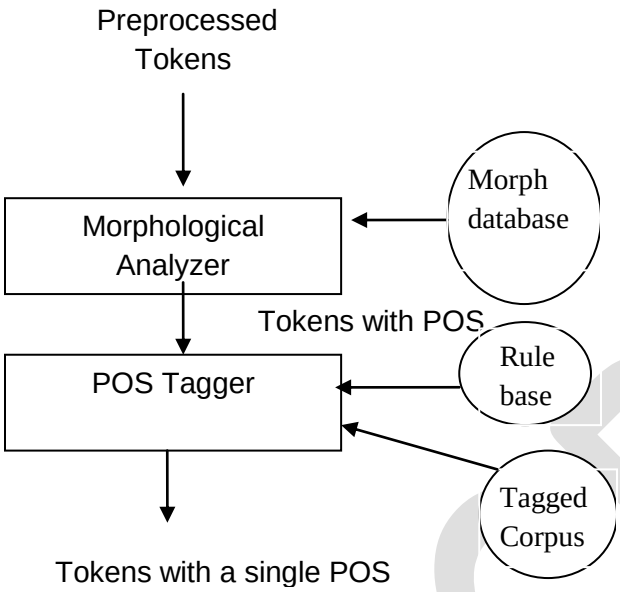
T: ieh merian kitaban han

E: these are my books

ieh (ieh) has two tags, i.e. showing singular and plural and according to the number attribute of auxiliary verb, it is translated to 'this' or 'these'.

3. Similarly the ambiguity related with the gender is resolved by considering the gender for surrounding words.

Fig 1 Morph Analyzer and POS Tagger



E.g.

P: auh kMm qy jw irhw sI

T: auh kam te ja riha see

P : auh kMm qy jw rhI sI

A. **T: auh kam te ja rahi see**

B. **Here** 'auh' **(auh) is translated to 'he' in first sentence and 'she' in second sentence depending upon the gender of the verb phrase.**

IV. EVALUATION OF TAGGER

The tagger was trained with 7204 tokens and then tested with 3426 tokens, from which 293 tokens came out to be unknown and tokens which were correctly tagged were 2867. The accuracy of the tagger came out to be 91.50%. The accuracy of the tagger can be further increased by training the system with more tagged tokens.

REFERENCES

1. Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In Third Conference on Applied Natural Language Processing (ANLP-92), pages 133--140, 1992

2. Church, Kenneth Ward. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In Proceedings of Second Conference on Applied Natural Language Processing, pages 136-143, Austin, Texas.

3. Cutting, Doug, Julian Kupiec, Jan Pederson, and Penelope Sibun. 1992. A Practical Part-of-Speech Tagger. In Proceedings of the Third Conference on Applied Natural Language Processing (ANLP-92), pages 133-140, Trento, Italy.

4. Brill, Eric. 1992. A Simple Rule-based Part-of-Speech Tagger. In Proceedings of the Workshop on Speech and Natural Language, pages 112-116, San Mateo, CA.

5. Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19(2):313-330.

6. Schmid, Helmut. 1994b. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the First International Conference on New Methods in Natural Language Processing (NemLap-94), pages 44-49, Manchester, UK.

7. Chanod, Jean-Pierre and Pasi Tapanainen. 1995b. Creating a Tagset, Lexicon and Guesser for a French Tagger. In Proceedings of the European Chapter of the ACL SIGDAT Workshop "From Text to Tags: Issues in Multilingual Languag Analysis", pages 58-64, Dublin, Ireland

8. Tlili-Guiassa Yamina, Tagging by Combining Rules- Based Method and Memory-Based Learning, World Academy of Science, Engineering and Technology 6 2005 pp 110-114

9. Yahya O. Mohamed Elhadj, Statistical Part-of-Speech Tagger for Traditional Arabic Texts, Journal of Computer Science 5 (11): 794-800, 2009

10. S. Bandyopadhyay and A. Ekbal. 2007. "HMM Based POS Tagger and Rule-Based Chunker for Bengali". In proceedings of the 6th International Conference on Advances on Pattern Recognition (ICAPR 2007), 2-4 January 2007, ISI, Kolkata, India, PP.384-390, World Scientific Press (Singapore) .

11. Brill, Eric. 1994. Some Advances in Transformation-Based Part-of-Speech Tagging. In Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), Vol.1, pages 722-727, Seattle, WA.

12. Brill, Eric. 1995b. Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging. In Proceedings of the Third Workshop on Very Large Corpora, pages 1-13, Massachusetts Institute of Technology, Cambridge, MA.

13. V. Goyal and G. S. Lehal, "Hindi Morphological Analyzer and Generator", Proceedings First International Conference on Emerging Trends in Engineering and Technology, Nagpur, G.H.Raisoni College of Engineering, Nagpur, July16-19, 2008, pp. 1156-1159, IEEE Computer Society Press, California, USA (2008)

14. Lezius, W.; Rapp, R.; Wettler, M. (1996). A morphology system and part-of-speech tagger for German. In: D. Gibbon (ed.): Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference, Bielefeld. Berlin: Mouton de Gruyter. 369-378.