ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/100213, August 2016* ISSN (Online) : 0974-5645

Effect of Statistical POS Tagger on Syntactic Analysis of Punjabi Sentences

Sanjeev Kumar Sharma*

Department of Computer Science and Applications, DAV University, Jalandhar – 144012, Punjab, India; Sanju3916@rediffmail.com

Abstract

Objectives: In this research article, author has explored the effect of statistics based part of speech tagger on the syntactic analysis of Punjabi sentences. **Methods/Statistical Analysis:** To study the effect of statistical POS tagger on the syntactic analysis of Punjabi sentence, author performed two experiments; first a rule based POS tagger is used for syntactic analysis and second this rule based POS tagger is replaced with HMM based statistical POS tagger. An annotated corpus of 20,000 words has been used to train the HMM based POS tagger. **Findings:** The system is tested on three types of errors; first subject/object and verb agreement error second noun and modifier agreement (in attributed form) error and third modifier and noun agreement error. On using HMM based POS tagger, the system shows a precision of 80.67 for subject/object and verb agreement error whereas on using rule based POS tagger the system shows a precision of 72.81. Similarly for noun and modifier agreement (in attributed form) error, author claims a precision of 82.45 on using HMM based tagger whereas on using rule based tagger, the precision is 76.00. And in case of modifier and noun agreement error, a precision of 97.56 is claimed by the author by using HMM based tagger which was 95.45 when rule based POS tagger is used. **Application/ Improvements:** The result indicates that the grammar checker performs better when rule based POS tagger is replaced with statistics based POS tagger.

Keywords: Punjabi Sentences, POS Tagger, Syntactic Analysis

1. Introduction

Syntactic analysis or grammar checking is one of the essential proofing tools for any natural language written by human being[1-2]. Grammar checking is also used as pre-processing and post-processing tool for some other natural language applications like machine translation. The Natural Language Processing is relatively new in Punjabi with some tools like machine translation (Hindi to Punjabi, Punjabi to Hindi, Gurumukhi to Shahmukhi, Shahmukhi to Gurumukhi), summarization system, OCR, grammar checker, spell checker etc. have been developed but a lot of work is going on to develop advance tools. Syntactic analyzer or grammar checker is one of the developed tools. A rule based grammar checker for Punjabi language has been developed[3-5]. This grammar checker uses a full form lexicon based morphological analyzer, a rule based POS tagger and a rule based phrase chunker as the essential component. Overall accuracy of grammar checker depends upon the accuracy of each of these components. In this research effect of POS tagger on the accuracy of grammar checker is evaluated. In subsequent sections of this research paper, role of POS tagger in grammar checking (Section 3), drawbacks of rule based POS tagger (Section 4) and statistics based POS tagger (Section 5) has been described.

2. Part of Speech Tagger

The main job of the POS tagger is to remove the ambiguity of tags that arises due to assignment of multiple tags to a word by the morphological analyzer. POS tagger is an important component of grammar checker. A word can exist in more than one form in different context like a noun can act as verb in some different context; similarly a verb can act as noun in some specific context. Therefore when input text is passed though morphological analyzer then morphological analyzer assigns all possible tags to

* *Author for correspondence*

each word and therefore, most of the words are assigned with two or more than two tags. This is the task of the part of speech tagger to select appropriate tag from these multiple assigned tags. A rule based POS tagger is developed for rule based grammar checker. The tagset used contains more than 630 tags[6]. The rules were developed by linguistic. The basic architecture of this rule based POS tagger is shown in Figure 1.

3. Role of POS Tagger in Punjabi Grammar Checker

As mention above, a rule based grammar checker has been developed[5]. This grammar checker checks the grammar on the basis of agreement check at phrase and clause level. This agreement is checked by using grammatical information provided by part of speech tags associated with each word. This grammar checker is composed of different components. These components includes pre-processor, morphological analyzer, part of speech tagger, phrase chunker and error detection/correction component working in a sequence i.e. output of one component is input to the other component. The basic architecture of this grammar checker is shown in Figure 2.

Figure 2 shows that part of speech tagger comes into role after the morphological analyzer. The morphological analyzer assigns all possible tags to each input word and it is the task of the POS tagger to assign the appropriate tag out of assigned tags to each word. Role of POS tagger

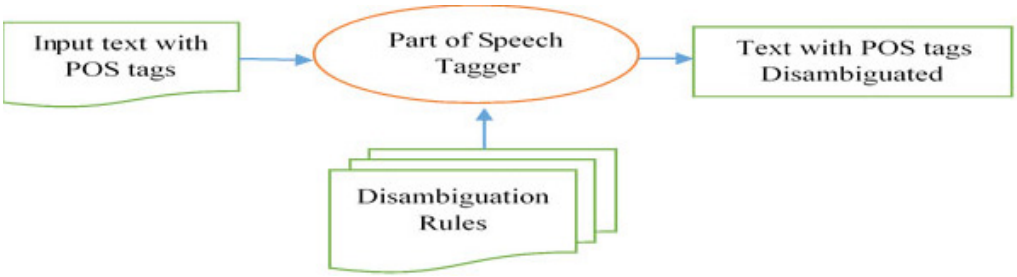

Figure 1. Basic architecture of rule based POS tagger.

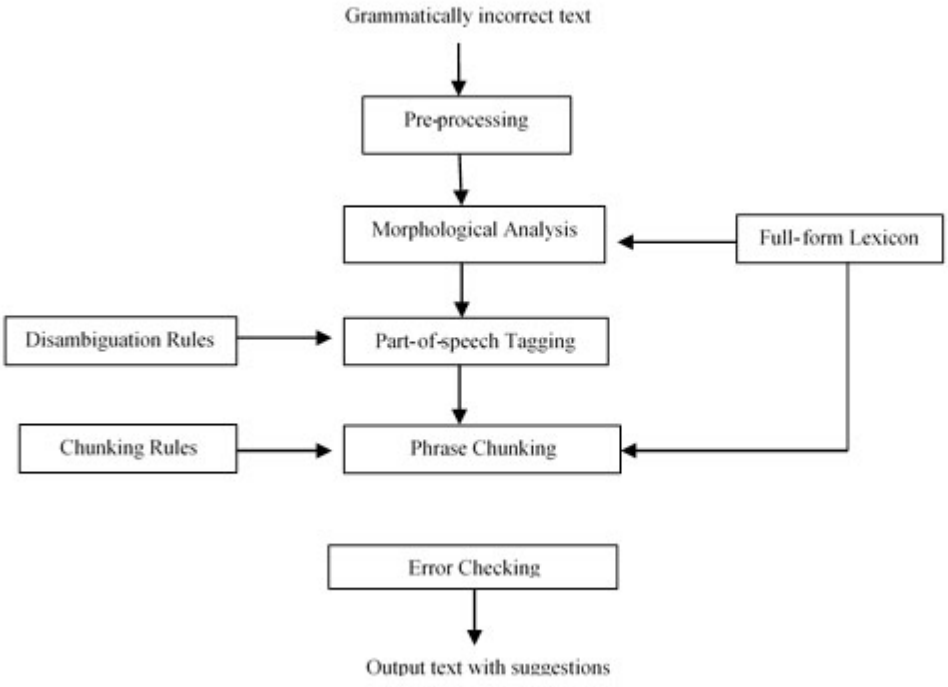

Figure 2. Basic architecture of Grammar checker.

is very essential for grammar checker point of view because the complete grammar checking depends upon these assigned part of speech tags. Consider the following example:

Punjabi: ਇਨਾਂਨੇਹਰਖੇਤਰਵਚਿਮੱਲਾਂਮਾਰੀਆਂਹਨ।

Transliteration: inhāṃnēharkhētarvicmallāṃmārīāṃ han.

English: These laurels in every area.

After passing through morphological analyzer, the output will be:

ਇਨਾਂ_PNDBPO ਨੇ_PPU ਹਰ_ NNMSD|NNMSO|AJUਖੇਤਰ_ NNMSD|NNMSO|NNMPDਵਚਿ_AVIBSD|PPIBSDਮੱਲਾਂ_ VBMAXSFXXTNE|NNBPD|NNBPOਮਾਰੀਆਂ_ VBMAFPXXPTNIA ਹਨ_VBAXBPT1

In above annotated (each word associated with part of speech tag) output there is four words which are assigned more than one part of speech tag by the morphological analyzer. These four words have been underlined. Out of these four words, three words (ਹਰ, ਖੇਤਰ and ਮੱਲਾਂ) have three tags associated to them whereas fourth word ਵਿੱਚ has two tags associated with it. Therefore, there are 3x3x3x2=54 possible combinations of the sentence and only one of them will be the correct. This is the job of the POS tagger to assign the correct tag to each word. Any wrong choice may lead to raise false alarm or incorrect grammar checking. Now when this output of morphological analyzer is fed to the rule based POS tagger, the outcome is:

(ਇਹਾਂ_PNDBPO ਨੇ_PPUNE ਹਰ_AJU ਖੇਤਰ_NNMSO ਵਚਿ_PPIBSD ਮੱਲਾਂ_VBMAXSFXXTNE ਮਾਰੀਆਂ_ VBMAFPXXPTNIAਹਨ_VBAXBPT1 ।_Sentence)

In above output the word ਮੱਲਾਂ has been assigned a verb tag (VBMAXSFXXTNE) instead of noun (NNBPD|NNBPO) tag. This result in the following incorrect output by the grammar checker:

Punjabi: ਇਨਾਂਨੇਹਰਖੇਤਰਵਚਿਮਾਰਆਿਮੱਲੇ।

transliteration: inhāṃnēharkhētarviccmāriāmallē

It is clear from above example that a wrong choice made by POS tagger results a false alarm. The rule based POS tagger used in this grammar checker is based upon the rules developed by linguistic.

4. Limitations of Rule based POS Tagger

The rule based POS tagger is composed of hand written rules developed by linguistic. As described by Naber (2003), it is not possible to develop an exhaustive set of rules for a language. Therefore very high accuracy cannot be obtained from a rule based system. The precision and

Table 1. Test results of rule based POS tagger

Corpus Genre	Total words	Unknown words	Tagged words		
			Incorrect tag (C)	Correct unique tag (A)	Ambiguous (at least one tag correct) (B)
Short stories	5469	291	228	4635	315
Book chapter	6278	910	258	4898	212
Essay	1934	153	90	1576	115
Thesis summary	5669	587	368	4412	302
Stories	5656	472	274	4557	353

Table 2. Precision and recall of rule based POS tagger

Corpus Genre	Rule based system				
	A	B	C	Precision	Recall
Short stories	4635	315	228	0.953115	0.936364
Book chapter	4898	212	258	0.949961	0.958513
Essay	1576	115	90	0.945978	0.931993
Thesis summary	4412	302	368	0.923013	0.935936
Stories	4557	353	274	0.943283	0.928106

Table 3. Test results of HMM based POS tagger

Corpus Genre	Total words	Unknown words	Tagged words		
			Incorrect tag (C)	Correct unique tag (A)	Ambiguous (at least one tag correct) (B)
Short stories	5469	291	218	4960	0
Book chapter	6278	910	312	5056	0
Essay	1934	153	102	1679	0
Thesis summary	5669	587	320	4762	0
Stories	5656	472	387	4797	0

recall obtained from this rule based system is provided in Table 1 and Table 2.

5. Statistics based POS Tagger

HMM based POS tagger has been developed[8-9]. This POS tagger was developed by using Hidden Markov Model. An annotated corpus of 20,000 words was used to train the system. The complete architecture of the statistical based POS tagger is given in Figure 3.

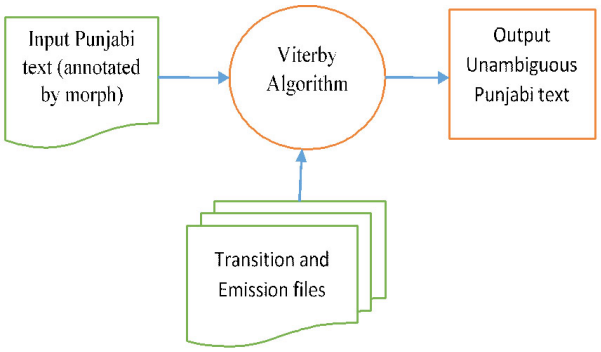

Figure 3. Basic architecture of HMM based POS tagger.

The precision and recall of this HMM based system is provided in Table 3 and Table 4.

Table 4. Precision and recall of HMM based POS tagger

Corpus Genre	HMM based system				
	A	B	C	Precision	Recall
Short stories	4960	0	218	0.957899	1
Book chapter	5056	0	312	0.941878	1
Essay	1679	0	102	0.942729	1
Thesis summary	4762	0	320	0.937033	1
Stories	4797	0	387	0.925347	1

6. Result and Discussion

Author has tested the system by taking three test sets. Each test set contains 50 sentences having a specific type of grammatical mistake. The three mistakes include Subject/Object and verb agreement error, Noun and modifier agreement error and Modifier and noun agreement error (in attributive form). These three sets are tested with both i.e. grammar checker having rule based POS tagger and

Table 5. Comparative analysis of rule based and HMM based system

Number of incorrect input sentences with error type	Output of the system			
	With Rule based tagger		With HMM based tagger	
	Precision	Recall	Precision	Recall
50 (Subject/Object and verb agreement error)	72.81	98.68	80.67	98.68
50 (Noun and modifier agreement (in attributive form)error)	76.00	76.00	82.45	78.19
50 (Modifier and noun agreement error)	95.45	100	97.56	100

grammar checker having HMM based POS tagger. When these test sets were syntactically analyzed after replacing the rule based POS tagger with this statistics based POS tagger a significant improvement was observed. The results obtained are shown in Table 5.

From results shown in Table 3, it is clear that there is significant improvement in the performance of grammar checker.

7. Conclusion

This technique can be further implemented to improve the performance grammar checking systems developed for other Indian and foreign languages. Further this approach can be used to improve other natural language applications where part of speech tagger plays an important role like sentence classification, speech processing etc.

8. References

1. Martin JH, Jurafsky D. Speech and Language Processing. International Edition. 2000.
2. Allen J. Natural Language Understanding. 1995.
3. Gill MS, Singh M. Development of a Punjabi grammar checker [PhD thesis]. Patiala: Punjabi University; 2008.
4. Gill MS, Lehal GS, Joshi SS. A punjabi grammar checker. IJCNLP. 2008. p. 940-4.
5. Gill MS, Lehal GS. A grammar checking system for Punjabi. In22nd International Conference on Computational Linguistics: Demonstration Papers; 2008. p. 149-52.
6. Gill MS, Lehal GS, Joshi SS. Part of speech tagging for grammar checking of Punjabi. The Linguistic Journal. 2009 May; 4(1):6-21.
7. Blunsom P. Hidden Markov Models. Technical Report. 2000
8. Sharma SK, Lehal GS. Using hidden Markov model to improve the accuracy of Punjabi POS tagger. IEEE International Conference Computer Science and Automation Engineering (CSAE); China. 2011. p. 697-701.