IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-9 E-ISSN: 2347-2693

Improving Existing Punjabi Morphological Analyzer using N-gram

S. K. Sharma

Dept. of Computer Science and Applications, DAV University, Jalandhar, India

[*]*Corresponding Author:* sanju3916@rediffmail.com

Available online at: www.ijcseonline.org

Received: 09/Aug/2017, Revised: 17/Aug/2017, Accepted: 14/Sep/2017, Published: 30/Sep/2017

***Abstract*--**Morphological analysis is an essential tool for almost all Natural Language Processes like POS tagging, Grammar checking, Sentence simplification, generation of Treebank and parsing. In this research article, author has used N-gram statistical technique to improve the existing morphological analyzer. The main factor that reduces the accuracy of morphological analyzer is presence of unknown words. In this research article author has used n-gram approach for detecting the POS tag of unknown word. The results shows an average precision of 82.34, recall 70.20 and F-measure 75.74.

***Keywords*--** Morphological analyzer, Morph, N-gram approach.

I. INTRODUCTION TO PUNJABI MORPHOLOGY

Punjabi, like other Indian languages, is morphologically rich language. It shows two types of morphology i.e. derivational morphology (adding prefix or suffix) and inflectional morphology (taking different form in different context). In derivational morphology, the word class of the word may change e.g. the word ਡਰ (ḍar) belongs to noun word class. But when author add a prefix ਨਿ to it, it becomes ਨਿ (ni) + ਡਰ (ḍar) = ਨਿਡਰ (niḍar), which is an adjective. Whereas in inflectional morphology, the word class of the word is preserved. For example, the inflectional forms of the word ਮੁੰਡਾ (muṇḍā) are ਮੁੰਡੇ (muṇḍē), ਮੁੰਡਿਆ (muṇḍiā), ਮੁੰਡਿਆਂ (muṇḍiāṃ), ਮੁੰਡਿਓ (muṇḍiō).

All these forms belong to noun word class but differ in number, gender and case. In Punjabi language, most of the words show inflection. This inflection results in creation of different morphological forms of a word and these different forms are marked with different part of speech tags.

II. MORPHOLOGICAL ANALYZER

It is a process or software that will take word as an input and return its root word along with other grammatical information related to this word. The grammatical information includes its number, gender, case and other applicable information as mentioned in table 1. General architecture of morphological analyzer has been shown in figure 1:

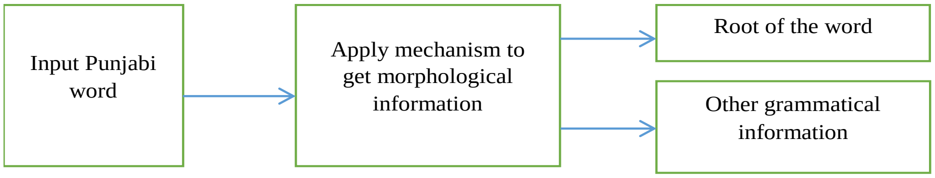
Input Punjabi word
Apply mechanism to get morphological information
Root of the word
Other grammatical information

Figure 1 : General architecture of morphological analyzer

As shown in figure 1, some mechanism is applied on input Punjabi word to obtain the root word and other grammatical information in the form of part of speech (POS) tags. For example, if the system takes input word ਸਾਨੂੰ (sānūṃ) then the morphological analyzer will return its root and part of speech tag as shown in table 2:

Table 1: Possible inflections shown by different word classes

Word Class	**Possible Inflections**
Noun	Number, gender and case.
Personal Pronoun	Number, gender, person and case
Reflexive Pronoun	Number, gender and case
Demonstrative Pronoun	Number, gender and case
Indefinite Pronoun	Number and case
Relative Pronoun	Number, gender and case
Interrogative Pronoun	Number, gender and case

Inflected Adjective	Number, gender and case
Cardinal	Case
Ordinal	Gender and case
Main Verb	Number, gender, person, phase, tense, transitivity, causality, inflectional classes.
Auxiliary Verb	Number, gender, person and tense
Inflected Adverb	Number, gender and case
Inflected Postposition	Number, gender and case
Vocative Particle	Number and gender
Conjunction, interjection, particle and verb-part.	Do not inflect

Table 2: Information provided by morphological analyzer

Input Word	**Root**	**Grammatical Information in form of Part of speech (POS) tag**
ਸਾਨੂੰ (sānūṃ)	ਮੈਂ (maiṃ)	PNPBPTF (personal pronoun inflected for both gender, plural number, first person)

III. EXISTING MORPHOLOGICAL ANALYZER:

A morphological analyzer has been developed by Gill and Lehal (2008). Full-form lexicon based approach has been used for its development. The drawback of this approach is that it needs a strong database. Although it is not possible to add all the words with their possible inflections including proper noun in the database, however, this database can be enriched by adding new words. Another problem with this approach is handling unknown words and similar words. Unknown words are those words that are not present in the database. These words generally arise due to spelling mistakes. In this research work, a thorough analysis of these types of words has been done and possible solutions have been proposed.

IV. THE TRIGRAM MODEL

It is observed that the word class of a word depends on the word class of previous two words. Therefore author calculated the trigram probability P(t3|t1, t2), where t3 stands for the word class of current word and t1 and t2 stand for the word class of two previous words. The word class for these previous two words are taken from the tagged training corpus. Similarly author can also calculate the word class of current word if author know the word class of previous and the next word to this current words. Another similar way is using the word class of next two words instead of previous two words. So author used all these three methods in following three different cases.

Case 1: If the POS tag of previous and next word to unknown is known to us, then author will calculate the trigram probability P (t3|t1, t2), where t3 stands for the unknown POS, and t1 and t2 stand for the previous and next word POS tags respectively.

Case 2: If the POS tag of previous word to unknown word is unknown which means previous word is also a unknown word, then author will calculate the trigram probability P (t3|t1, t2), where t3 stands for the unknown POS, and t1 and t2 stand for the POS tags of next two words.

Case 3: If the POS tag of next word to unknown word is unknown which means next word is also a unknown word, then author will calculate the trigram probability P (t3|t1, t2), where t3 stands for the unknown POS, and t1 and t2 stand for the POS tags of previous two words.

Now in order to calculate the trigram probability an annotated corpus was developed. This corpus was collected from different Punjabi authors sites by keeping in mind that all the common domains should be covered. Then this corpus was tagged by using a pre-existing rule based POS tagger. This pre-existing POS tagger uses 630 tags which covers almost all the word classes with their inflections. This trained POS tag was divided in to two different corpuses, one containing the sentences without any unknown word and the other containing the sentences that contain unknown words. The corpus that does not contain any unknown word was used for training the model and the other portion that contains unknown words was used for testing. The complete architecture is shown in figure 2.

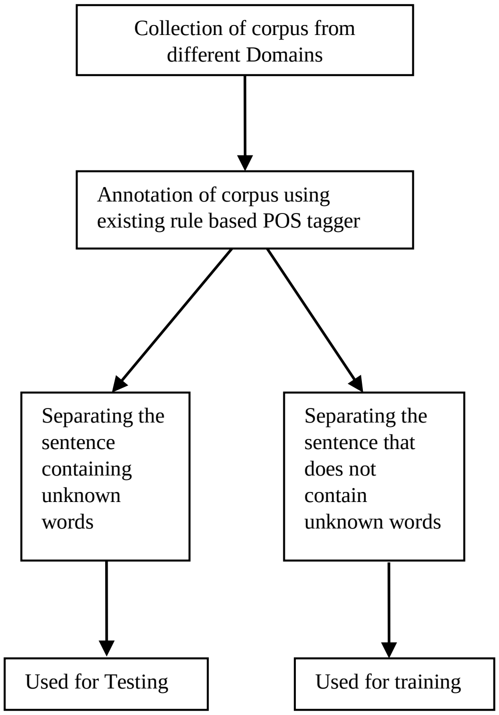
Collection of corpus from different Domains
Annotation of corpus using existing rule based POS tagger
Separating the sentence containing unknown words
Separating the sentence that does not contain unknown words
Used for Testing
Used for training

Figure 2: Architecture of proposed N-gram morphological analyzer

A. Collection of Corpus

The basic need for using the statistical methods/techniques is the availability of annotated corpus. More the corpus available more will be the accuracy. Another thing that should be kept in mind is that the corpus should be accurate. So author started our work with the collection of accurate corpus. While collecting the corpus author kept the following points in our mind:-

- The corpus should be in Unicode.
- The corpus should be accurate i.e. it should have minimum no of spelling mistake.
- The corpus should not be domain specific.
- The corpus should contain as many different words as possible.

The main sources of our corpus are:

- http://punjabikhabar.blogspot.com
- http://www.quamiekta.com
- http://www.europediawaz.com
- http://www.charhdikala.com
- http://punjabitribuneonline.com
- http://www.sadachannel.com
- www.veerpunjab.com
- www.punjabinfoline.com

B. Annotation of the corpus

Annotation of the corpus means giving a tag to the every individual word. The next step that author performed after the collection of corpus was to annotate the corpus. Author annotate the corpus by using a tool named TAGGER. This tool is developed by author from a pre-existing Rule based POS Tagger. Author made some alteration in that pre-existing tool and used it for the annotation of corpus.

C. Screening/Filtering of annotated corpus

As the annotated corpus contains many words having ambiguous tags i.e. the words having more than one tag, so author filtered the sentence that contains ambiguous words. In this way author divided the annotated corpus in two parts, one containing the sentences that have ambiguous words and the other that does not contain any sentence having ambiguous word. After this first filtering author applied another type of filtering. From the annotated corpus that does not contain any ambiguous word author separate the sentence that does not contain unknown words.

D. Creating Triplets with frequency

From the corpus that does not contain any unknown word author created the triplets of part of speech (POS) tags. After creating triplets, author calculate their frequency of occurrence in the corpus.

Table 3: Sample Triplet Table

Triplet	Frequency
NNFSD_VBP_VBMAXSS3XBNO	1
VBP_VBMAXSS3XBNO_PTUKE	1
VBMAXSS3XBNO_PTUKE_PNPMPGDF	2
PTUKE_PNPMPGDF_NNMSO	3
PNPMPGDF_NNMSO_PPIBSD	44
NNMSO_PPIBSD_NNMSD	119
PPIBSD_NNMSD_VBMAMSXXPINIA	9
NNMSD_VBMAMSXXPINIA_CJC	21
VBMAMSXXPINIA_CJC_AVU	8
AJIFSD NNFSD VBMAFSXXPTNIA	70
NNFSD VBMAFSXXPTNIA CJC	40
VBMAFSXXPTNIA CJC NNFSD	5
CJC NNFSD VBMAXSS3XBNO	3
CJC_AVU_PTUE	35

V. RESULTS AND DISCUSSION

Author divide the testing corpus into four parts of equal length. These four equal parts contains different number of unknown words. Results obtained by author are tabulated in table 4.1 and table 4.2.

Table 4.1: Results Obtained

Total words in the corpus	**No of unknown words A**	**Correctly tagged Unknown words B**	**Incorrectly tagged Unknown words C**	**Not tagged**
12430	547	392	92	63
12450	345	254	30	61
12444	355	225	25	105
12465	456	329	73	54

Table 4.2: Results Analysis

Sr. No.	**Precision** $\frac{B+C}{A}$ X 100	**Recall** $\frac{B}{A}$X 100	**F-Measure** $\frac{Precision\ X\ Recall}{precision+recall}$ X2
1	88.48	71.66	79.18
2	82.31	73.62	77.72
3	70.42	63.38	66.71
4	88.15	72.14	79.35
Average	82.34	70.20	75.74

As shown in table 4 and table 5, the proposed n-gram system shows an average precision of 82.34, Recall 70.20 and F-Measure 75.74. The reason for not tagging the unknown word is absence of the triplet with that combination. Most of the untagged unknown words are of similar type. The incorrect tags of unknown words can be further reduced by selecting two highest frequency triplets satisfying the condition. Suppose author have a word **ਵਿਲੇਨ** in following sentence:

ਇਸ ਫ਼ਿਲਮ ਦੇ ਅੰਤ ਵਿੱਚ ਵਿਲੇਨ ਨੂੰ ਪਿਸਤੌਲ ਨਾਲ਼ ਸ਼ੂਟ ਕਰਨ ਦੀ ਬਜਾਏ ਜ਼ਹਿਰ ਦੇ ਕੇ ਮਾਰਿਆ ਜਾਂਦਾ ਹੈ।

(ਇਸ_PNDBSO ਫ਼ਿਲਮ_NNFSO ਦੇ_PPIDAMSO ਅੰਤ_NNMSO ਵਿੱਚ_PPIBSD ਵਿਲੇਨ_Unknown ਨੂੰ_PPUNU ਪਿਸਤੌਲ_Unknown ਨਾਲ਼_AVU ਸ਼ੂਟ_NNFSD ਕਰਨ_NNMSO ਦੀ_PPIDAFSO ਬਜਾਏ_PPU ਜ਼ਹਿਰ_NNMSO ਦੇ_PPIDAMSO ਕੇ_PPIMPD

ਮਾਰਿਆ_VBMAMSXXPTNIA ਜਾਂਦਾ_VBOPMSXXXINDA ਹੈ_VBAXBST1 ।_Sentence)

In above sentence ਵਿਲੇਨ is unknown word.

When author search for the triplet
PPIBSD _Unknown _PPUNU
Author get many combinations with different frequencies but the two highest frequencies are
PPIBSD _NNMSO_PPUNU 54
PPIBSD _NNFSO_PPUNU 48
So instead of replacing Unknown with NNMSO (with highest frequency 54) author prefer replace Unknown with NNMSO/NNFSO. Further the POS tagger will resolve this ambiguity.

REFERENCES

[1]. Bharati, Akshar, Amba P. Kulkarni, Vineet Chaitanya. (1998a).*Challenges in Developing Word Analyzers for Indian Languages*, Presented at Workshop on Morphology, CIEFL, Hyderabad, July 1998.

[2]. Bharati, Akshar, Rajeev Sangal and S.M. Bendre (1998b). *Some Observations on Corpora of Some Indian Languages*. Knowledge Based Computing Systems, Tata McGraw-Hill.

[3]. Goldsmith, John. (2001). *Unsupervised Learning of the Morphology of a Natural Language*. Computational Linguistics, Vol 27, No. 2, pp 153-198.

[4]. Daniel Jurafsky, James H. Martin. *Speech and Language Processing:An introduction to speech recognition*, Natural Language Processing, and Computational Linguistics. LTRC, IIIT Hyderabad http://ltrc.iiit.ac.in

[5]. Gill Mandeep Singh, Lehal Gurpreet Singh, Joshi S.S., *A full form lexicon based Morphological Analysis and generation tool for Punjabi*, International Journal of Cybernatics and Informatics, Hyderabad, India,October 2007, pp. 38-47

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

[7]. Cutting, J. Kupiec, J. Pederson and P. Sibun, *A practical part of-speech tagger*. In Proc. of the 3rd Conference on Applied NLP, pp. 133-140, 1992

[8]. Dermatas and K. George, *Automatic stochastic tagging of natural language texts*. Computational Linguistics, 21(2): 137-163, 1995

[9]. Ekbal, Asif, and S. Bandyopadhyay,"*Lexicon Development and POS tagging using a Tagged Bengali News Corpus",* In Proc. of FLAIRS-2007, Florida, 261-263, 2007

[10]. E. Dermatas and K. George, *Automatic stochastic tagging of Natural language texts*, Computational Linguistics, 21(2): 137-163, 1995

[11]. Ekbal Asif, et.al, *"Bengali Part of Speech Tagging using Conditional Random Field"* in Proceedings of the 7th International Symposium of Natural Language Processing (SNLP-2007), Pattaya, Thailand, 15 December 2007, pp.131-136