Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 1): pp. 4200-4211
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.462

Assigning Correct Semantic Information Using Stochastic Model

Vikas Verma Research Scholar, DAV University, Jalandhar, India
vikas2005verma@yahoo.co.in.

S.K. Sharma DAV University, Jalandhar, India sanju3916@rediffmail.com

Abstract:

In this era of machine learning and artificial intelligence, natural language processing applications are in high demand. For the development of almost all the natural language processing applications assignment of semantic informationis the fundamental component. This component assign the correct semantic information to individual word. In this research paper,we proposea stochastic approach for the development of system to assign correct semantic information for one of the morphologically rich Indian language i.e. Punjabi language. We have used annotated corpus available by Indian language corpora initiative to train the system. Thisdata is used for training and approximately 99k sentences collected form internet are used for testing.Further the developed system is also capable to assign the part of speech tag to unknown words. On testing the system, author claimed an accuracy of 89.33%.

Keywords: Punjabi POS tagger, Stochastic, NLP, Grammar Checker.

Introduction:

With the advancement of artificial intelligence, natural language processing applications are in hot demand. Whether we talk about text processing systems, machine translation systems, text to speech system, speech to text systems, chat boats and many more, all are directly or indirectly applications of natural language processing. Siri, Alexa and Google Assistant are some of most commonly used software. For the development of almost all the natural language processing applications some fundamental task are required. Part of speech tagger is one of such task that is performed as pre-process for almost all the NLP applications. In every language, there are many words which plays different roles in different context. When morph is applied on such word then morph will assign all possible part of speech tags that this word can have. The fundamental task of the part of speech tagger is to disambiguate the multiple part of speech assigned to single word.This task

becomes more difficult if the language is morphologically rich and almost all Indian languages are morphologically very rich. Therefore, it becomes a challenging taskto develop a part of speech tagger for Indian languages. This attracts a number of researchers to develop the POS tagger for Indian languages.

Introduction to Part of speech tagger:

Tagging can be considered as a classification problem in which one has to assign the correct word class (nouns, verb, adverbs, adjectives, pronouns, conjunction and their sub-categories) to the token. The word class is in the form of tag called part of speech tag or semantic information. Consider the following Punjabi example:

Punjabi: ਦੋ ਸੋਹਣੇ ਮੁੰਡੇ ਜਾਂਦੇ ਹਨ
Transliteration:(do sohane mude jande han)

Translation:Two handsome boys go.
After assigning semantic information the output will be:
ਦੋ_ CDPDਸੋਹਣੇ_AJUਮੁੰਡੇ_NNMPD ਜਾਂਦੇ_ VBMAMPXXXINDA ਹਨ_ VBAXBST1
In above example, the grammatical information is assigned to each word in the form of tags called part of speech (POS) tags

Introduction to Punjabi language:

Punjabi is one such morphologically rich Indian language. It belongs to Indo-Aryan family of languages. It is mostly used in the north regions of India and West Pakistan. It is spoken as the primary language in the states of Punjab. It is spoken by more than 100 million persons in India. Other than India, Punjabi language is spoken by many migrated Indian in Canada, UK, Australia and many other countries.

Existing Work:

Basic approaches used for development of POS tagger falls into two categories i.e. stochastic based approach and rule based approach. Stochastic based approach is used when the developer has very small knowledge of the language for which POS tagger is to be developed. Also this approach is easy to implement. On the other hand to implement rule based approach, thorough knowledge of language is required. This approach is difficult to implement as compare to stochastic based approach.A number of authors tried different approaches to develop POS tagger for different languages. These includes POS tagger for English language using transformation rules by Brill [1]. The system showed an accuracy of 95 %. Another POS tagger for Myanmar using HMM was developed by Zin and Thein [2] and reported 97.56 % accuracy. Like in Hindi Language also, many researchers tried to develop POS tagger. These include AnnCorra (Lexical Resources for Indian Languages

(LERIL) has a project called "Annotated Corpora" [3]).Further, Mishra and Mishra [4] used rule based approach, Garg et al. [5] also implemented a POS system with 85.47% precision using rule-based approach, Singh et al. [6] used morphological analysis and CN2 algorithm and got 93.45 % of accuracy, Dalal et al. [7] used maximum entropy Markov model and obtained 94.38 % accuracy.Various techniques have been used by various researchers to develop POS tagger for Punjabi language (Language similar to Hindi language). These techniques includes Rule based approach used by [9], HMM based approach used by [10], N-gram based technique used by [11], Reduced tagset used by [12], Neural network based POS tagger developed by [13], GA based approach used by [14], bi-gram based approach used by [15], hybrid approach i.e. combination of HMM and rule based by [16],SVM based technique used by [17] and machine learning approaches used by [18].

Proposed approach:

In this research, author used stochastic technique to develop the Punjabi POS tagger. This technique is a statistical techniques and is widely used to develop natural language processing applications like detection of unknown POS tag in Punjabi language [15], to develop part of speech tagger for Punjabi language[11][13], to develop grammar checker [21][22][23][24]. To implement this technique, a large amount of annotated corpus is required. This annotated corpus can be either manually generated or standard annotated corpus available for research. In this research, author used pre-existing standard annotated corpus available at Indian Languages Corpora Initiative [20]. This research work is completed in two phases. In the first phase, trigrams probabilities from annotated corpus are calculated and stored in a database. In the second phase, input text is assigned appropriate parts of speech as per the trigram probabilities calculated in first phase. In the following section, both of these phases are discussed briefly.

Phase 1 (Creation of trigram probabilities):

Annotated corpus available at ILCI is used to create trigram probabilities. First all the ILCI corpus is broken into sentences and then from each sentence part of speech tags are extracted. From these extracted POS tags, trigrams are generated. This can be further explained by following example:

Annotated Sentence: ਦੇ_ CDPDਸੋਹਣੇ_AJUਮੁੰਡੇ_NNMPD ਜਾਂਦੇ_ VBMAMPXXXINDA ਹਨ_ VBAXBST1

Extracted POS pattern: CDPD_AJU_NNMPD_ VBMAMPXXXINDA_ VBAXBST1

Trigrams of POS patterns:

Following trigrams are generated from the POS patterns

CDPD_AJU_NNMPD
AJU_NNMPD_ VBMAMPXXXINDA
NNMPD_ VBMAMPXXXINDA_ VBAXBST1
VBMAMPXXXINDA_ VBAXBST1_None
VBAXBST1_None_None

Probability to calculate trigram = $\frac{\text{Count (ti−1_ti_ti+1)}}{\text{Total number of trigrams}}$

After generation of trigrams, probability of each trigram is calculate by using the following formula:

In above formula t_i is the current tag, t_{i-1} is the previous tag and t_{i+1} is the next tag. Some sample entries are provided in the following table 1:

Table1: Sample entries of trigram probabilities

Trigrams	Probability
NNFSD_VBP_VBMAXSS3XBNO	0.124065
VBP_VBMAXSS3XBNO_PTUKE	0.110202
VBMAXSS3XBNO_PTUKE_PNPMPGDF	0.038351
PTUKE_PNPMPGDF_NNMSO	0.143304
PNPMPGDF_NNMSO_PPIBSD	0.219112
NNMSO_PPIBSD_NNMSD	0.733298
PPIBSD_NNMSD_VBMAMSXXPINIA	0.188384
NNMSD_VBMAMSXXPINIA_CJC	0.059797
VBMAMSXXPINIA_CJC_AVU	0.128332
CJC_AVU_PTUE	0.14809
AVU_PTUE_NNFSO	0.039182
PTUE_NNFSO_PPIDAFPD	0.392429
NNFSO_PPIDAFPD_NNFSD	0.089174
PPIDAFPD_NNFSD_VBMAFPT3XINEGA	0.03885
NNFSD_VBMAFPT3XINEGA_Sentence	0.06919

The complete work flow to generate the trigram probability is shown in

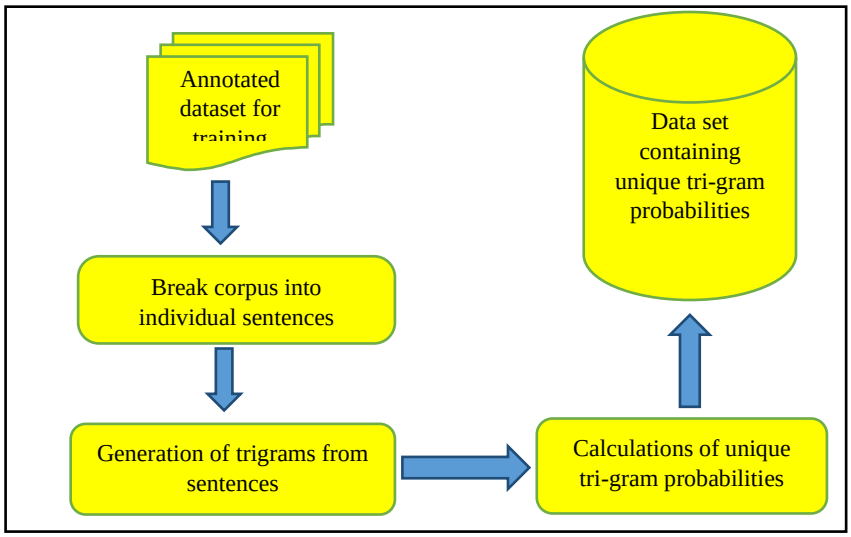

figure1.

Figure 1: Generation of trigrams from annotated corpus

Phase 2: (Assigning correct part of speech tag to each token):

In this phase, correct part of speech tag is assigned to each tag using the trigram probabilities calculated in phase 1. As in each language there are many words which have different semantic information in different context. Consider the following example:

ਸਿਪਾਹੀਦੀਮੌਤਹੋਗਈ_sipahi di mauta ho ga'i (Soldiers die)

ਮ੍ਰਿਤਕਸਿਪਾਹੀਨੂੰਬਚਾਇਆਗਿਆ_mritaka sipahi nu baca'i'a gi'a (Dying soldier rescued)

ਜੰਗਵਿੱਚਫੌਜੀਮਰਦੇਨਹੀਂਸ਼ਹੀਦਹੁੰਦੇਹਨ।_jaga vica phauji marade nahim sahida hude hana. (In war, soldiers do not die but martyrs)

In above three sentences, the word ਮਰਦੇ(marade) has different semantic meaning. In the first sentence, word ਮਰਦੇ(marade) is verb, in second sentence word ਮਰਦੇ(marade) is adjective and in third sentence word ਮਰਦੇ(marade) is noun. When morphological analyzer is applied on any of the above sentences then the three different types of grammatical information will be assigned to word ਮਰਦੇ(marade). Thus the results obtained after applying morphological analyzer will be:

(ਸਿਪਾਹੀ_NNMXD ਮਰਦੇ_VBMAMPXXXINDA/AJIMSO/NNMXDਗਈ_VBOPMPXXPINIA ।_Sentence)

(ਮਰਦੇ_VBMAMPXXXINDA/AJIMSO/NNMXDਸਿਪਾਹੀ_NNMSO ਨੂੰ_PPUNU ਬਚਾਇਆ_VBMAMSXXPTNIA ਗਿਆ_VBOPMSXXPINIA ।_Sentence)

(ਜੰਗ_NNFSO ਵਿੱਚ _PPIBSD ਸਿਪਾਹੀ_NNMXD ਮਰਦੇ_ VBMAMPXXXINDA/AJIMSO/NNMXDਨਹੀਂ_PTUN ਸ਼ਹੀਦ_AJIBXD ਹੁੰਦੇ_VBMAMPXXXINDA ਹਨ _VBAXBPT1 ।_Sentence)

As shown above, the word ਮਰਦੇhas been assigned three types of grammatical information i.e. VBMAMPXXXINDA/AJIMSO/NNMXD. Now the POS tagger's job is to assign the tags which are correct POS tag. As per our algorithm this problem will be solved as follow:

First trigrams of the input sentence having ambiguous words (multiple tags to a specific word) are generated. Therefore in the first sentence, the trigrams will be:

NNMXD _VBMAMPXXXINDA _VBOPMPXXPINIA

NNMXD_AJIMSO _VBOPMPXXPINIA

NNMXD_NNMXD _VBOPMPXXPINIA

Probability of all above trigrams is checked from the trigram probability database created in phase1. The trigram having maximum probability will be assigned as the correct POS tag. The complete architecture of this phase is explained in figure 2.

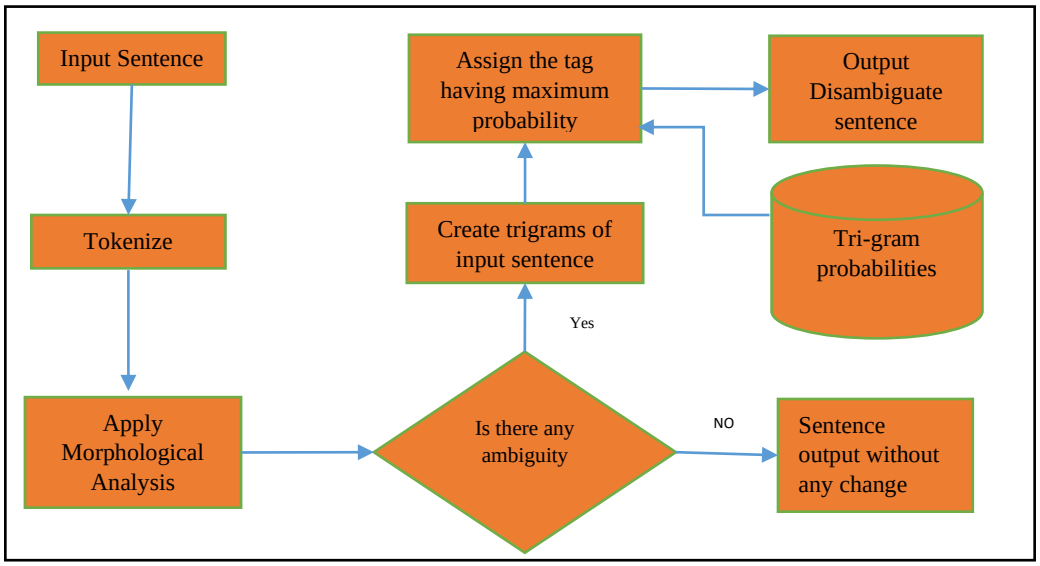

Figure2: Calculation of trigram probabilities

Annotated data used for creating trigrams:

The annotated corpus used for creation of tri-grams is taken from Indian Languages Corpora Initiative (ILCI). This data is freely available at [20].for research purpose. The details of the annotated corpus is provided in table 2.

Table 2: Details of corpus taken from ILCI

Details of Domain in the corpus	Total files	Overall sentences count (approximate in thousands)	Overall words count (approximate in thousands)
Agriculture	15	120	1023
Entertainment	15	137	1432
Tourism	15	114	1209
Health	15	121	1378
Total	60	492	5042

Size of training set:492Ksentences (approximate)

Size of testing set: 99K sentences (approximate) collected from reliable online resources like newspaper websites.

Result and discussion:

After calculating the tri-gram probabilities from training set, testing of the developed system is performed using testing data. The testing data is divided into five sets (approximate 20000 sentences in each set). On testing the system, the results obtained are tabulated in table 3.

Table 3: The outcomes of the proposed system's testing

Set No.	Number of sentences (approx in thousand)	No. of words (approx in thousand)	No. of words having ambiguity (approx in thousand) (A)	Number of words assigned correct tag by system (B)	Number of words assigned incorrect tag (C)	Precision $\frac{B+C}{A}$ X 100	Recall $\frac{B}{A}$X 100	F score $\frac{\text{Precision X Recall}}{\text{precision+recall}}$ X 2
Set_1	20	182	21	18845	291	91.12	89.74	90.43
Set_2	20	206	14	12391	383	91.24	88.51	89.85
Set_3	20	211	25	22458	321	91.12	89.83	90.47
Set_4	20	199	19	16895	433	91.20	88.92	90.05
Set_5	20	223	32	28573	601	91.17	89.29	90.22
Total	100	1021	111	99162	2029	91.16	89.34	90.24

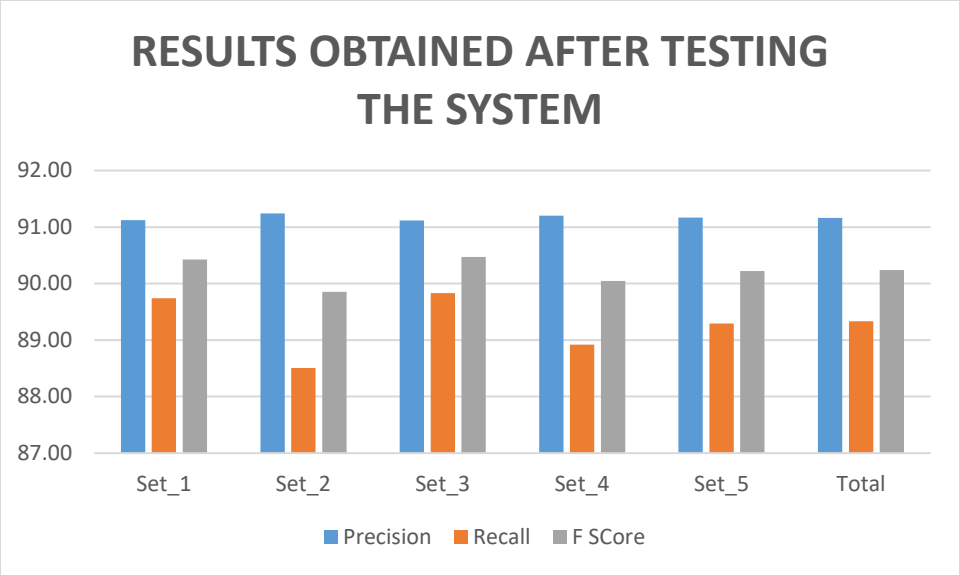

Figure 3: Results obtained on testing the proposed system

Further overall accuracy is calculate using the following formula:

$$\text{Overall accuracy} = \frac{\text{Number of correctly Tagged Words}}{\text{Total Number of ambigous words}}$$

$$= \frac{99162}{111000} = 0.8933 \text{ or } 89.33\%$$

Comparative analysis with existing systems:

Author compared the developed system with state of art existing systems. The developed system was compared with HMM with Reduced tagsetsystem developed by Manjit kaur et al. [12], System developed using Genetic Algorithm by Kamaljot Singh [14], Support Vector System developed by Kumar and Josan [17], HMM and Rule based System developed by Singh and Goyal [25]. The comparative analysis is shown in table 4:

Table 4: Comparison with existing approaches

POS tagger systems	Technique used	Accuracy
Manjit kaur et al. [12]	HMM with Reduced tagset	93.5
Kamaljot Singh [14]	GA	90.63
Kumar and Josan [17]	SVM	89.86
Singh and Goyal [25]	HMM and Rule based	93

Proposed system	Stochastic technique	89.33

Figure 4: comparison of proposed system with existing systems

Conclusion and Future Scope:

In this research paper, author proposed stochastic part of speech tagger for Punjabi language (one of the most morphologically rich Indian languages). Author used trigram probability to disambiguate the ambiguous tags. To create the trigram probability author used annotated corpus provided by Indian language corpora Initiative (ILCI). This whole corpus is used to develop trigram probabilities and approximate99k sentences collected from reliable online resources are used to test the system. On testing the system on this 99k sentence corpus, author obtained an accuracy of 89.33%. From result it is observed that the proposed system achieved comparable accuracy as compare to other state of art developed POS taggers. One more advantage of author's proposed system is that it can also assign the grammatical information to the unknown words i.e. word which is not present in the lexicon. In future, this accuracy can be further enhanced by increasing the training data and also the developed technique can be applied on other similar languages (morphological rich languages).

References

[1]. Brill, E.: A simple part of speech tagger based on rules. In Proc. of the 3rd Conference on ANLP , Stroudsburg, PA, US, pp. 152–155 (1992)

[2]. Zin, K.K., Thein, N.L.: POS tagging for Myanmar using HMM. In Proc. of ICCTIT, Dubai, pp. 1–6 (2009)
[3]. Sharma, D.M., Bharati, A., and Sangal, R.: An Introduction AnnCorra (Vol. 14), Technical Report : TR-LTRC (2001)
[4]. Mishra, A., Mishra, N.: POS for Hindi corpus. In Proc. of the ICCSNT, India, pp. 554–558 (2011)
[5]. Garg, N., Preet, S, Goyal, V.: POS Hindi tagger based on rules. In Proc. Of Coling, India (2012)
[6]. Singh, S., Shrivastava, M., Gupta, K., Bhattacharyya, P.: Resource poverty richness Morphological offsets—an experience building Parts of speech tagger for Hindi. In Proc. of Coling, Australia (2006)
[7]. Dalal, A., Sawant, U., Nagaraj, K., Bhattacharyya, P., Shelke, S.: Building POS tagger with rich feature for morphologically rich languages: Hindi experience. In Proc. ICON (2007)
[8]. A POStagger for Indian languages (POS tagger) (2007)
[9]. Gill, M. S., Joshi, S. S., Lehal, G. S.: POS tagging forPunjabi grammar checking. The Linguistic Journal, 4(1), 6-21.
[10]. Sharma, S. K., & Lehal, G. S. Using HMM to improve the accuracy of POS tagger in Punjabi. In CSAE, IEEE International Conference ,vol.2, 697-701(2011).
[11]. Mittal, S., Sharma, S. K., Sethi, N. S.: POS tagging in Punjabi using N-Gram Model. IJCA, 100(19) (2014).
[12]. Kaur, M., Aggarwal, M., & Sharma, S. K.: Improving POS tagging in PunjabiUsing Reduced Tag Set. IJCA, 7(2), 142 (2014).
[13]. Kashyap, D. K., & Josan, G. S.: A trigram model to predict POS tags using NN (Neural network). In INIDEAL, Springer, Berlin, Heidelberg.(pp. 513-520)(2013).
[14]. Singh, K. (2015). POS Tagging using GA (Genetic Algorithms). In IJSSST, 16(6).
[15]. Sood, S., Sharma, S. K., Arora, V.: Word Class Prediction of Unknown and Ambiguous Words of Punjabi using Bi-gram Methods. In IJCAIT, 7(2), 152.
[16]. Kanwar S.,Ravishankar, Sharma, S.K.: POS tagging of Punjabi language Using HMM. In IJES. pp 98-106. (2011).
[17]. Kumar, D., & Josan, G. POS tags Prediction in Punjabi using SVM (Support Vector Machines). In IAJIT (International Arab Journal of Information Technology), 13(6) (2016).
[18]. Kumar D. and Josan G.: Developing amachine learning tagset for POS tagging in Punjabi language. In IJARITC,vol. 3(2), pp. 132-143. (2012).
[19]. Deepa M., Neeta N.: POS Tagging Hindi Corpus Using Rule-Based Method. InProc. of the ICRCWCIP. Springer, New Delhi.(2016).
[20]. http://sanskrit.jnu.ac.in/ilci/index.jsp. Accessed on Oct 9, 2021.

[21]. Alam, M., UzZaman, N., and Khan, M.: N-gram based Bangla and English statistical grammar checker.
[22]. Nazar, R., & Renau, I.:N-Gram Google based corpus used as a grammar checker. In Proc. of Second Workshop on CLW. Linguistic and Cognitive Aspects of DCDE (Document Creation and Document Engineering), ACL. pp. 27-34.(2012)
[23]. Temesgen, A., & Assabie, Y.: Development using morphological features for Amharic grammar checker on words and N-gram based probabilistic methods. In Proc. of 13th ICPT (International Conference on Parsing Technologies). pp. 106-112.(2013)
[24]. Go, M. P., & Borra, A.: Developing for Filipino an unsupervised grammar checker using hybrid N-grams as grammar rules. In Proc. of 30th PACLIC (Pacific Asia Conference on Language, Information (2016).
[25]. Singh, Umrinder and Goyal, Vishal.: Punjabi POS tagger: Rule Based and HMM. International journal of computer science and software Engineering. (2017).