www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 3 March 2017, Page No. 20780-20784
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i3.68

Automatic Rule Detection and POS Tagging of Punjabi Text

Er.Davinder Kaur[1], Er.Ubeeka Jain[2]
[1] Research Scholar
[2]Faculty of Computer science and Information Technology, I.K. Gujral Punjab Technical University, Rayat Institute of Engineering and Information Technology, India
[1]davinder.uppal@yahoo.in ,[2]ubeekajain@gmail.com

Abstract-Natural language is a study that involves the communication between humans and computers. There are number of tasks that can be performed by using NLP like speech recognition, information extraction, segmentation, language translation, grammer checking etc.. Punjabi language is mostly spoken by the people living in Punjab and Pakistan. NLP is considered as a hard problem in computer science due to the ambiguity of the languages .POS tagging involves process for disambiguating the part-of-speech information for such ambiguous words by taking into consideration the context information. In this paper we study this problem and generate " automatic rule detection and pos tagging of Punjabi text" by using rule based approach. The chances of error are less and work is more efficient and accurate.

Keywords – Punjabi language, taggers, pos tagging

Introduction

A part-of-speech tagger is a system that uses context to assign parts of speech to words. Taggers may use several kinds of information like dictionaries ,rules etc. Taggers mainly used two kind of approaches that is rule based and statistical approach. In rule based approach knowledge base is developed by an expert where to assign the pos tag. In the statistical approach language models are built and used to pos tag the input text automatically. Words in punjabi language often consist of concatenation of word segments,each corresponding to a part of speech(POS) category.There may be many words that can have more than one tag. The study goals were 1) to develop annotated pos data. 2) to develop an algorithim to extract rules from annotated text.3) to develop rule based pos algorthim. In this system we used rule based approach which uses the training data and produces an inferred rule that can be used for creating new examples ,each example consists of an input object and required output object. In this system whenever Punjabi text is input by the person ,firstly it performs tokenization and normalization on that input string . Afterthat it try to find out tags from the database , if the tag is not found in the database then it apply bigram rules to tag that word .This system assign two types of tags that is unigram(words having one tag) and bigram(words having two tags).

We have studied different research papers related to Part of Speech Taggers which are based on different approaches. Most of the taggers developed using different statistical approaches and those systems are only specific to UNIX based operating system. So there are very few Part of Speech taggers available for Windows operating system. We have developed Part of Speech tagger for Punjabi using rule based approach. Where we have developed algorithm to extracted rules from training data automatically.

This system was developed using 30 different standard part of speech tags that are given by Department of Information Technology Ministry of Communications & Information Technology and some other tags that is time, date and number tag. Adverb tag is further classified in to following categories i.e.

Adverb of manner (RB_AMN)

Adverb of location (RB_ALC)

Adverb of time (RB_TIME)

Adverb of quantity (RB_Q)

Collection of 15,114 words for different tags has been done. The system mainly works in two steps-firstly the input word is match with the database; if it is present then it is tagged. Secondly input word is not president in one to one mapping database then system used extracted rules to tag new words.

We have used n-gram model to extract rules from training data. Where value of n is two. n-gram model can be defined as:

$$Bi - Gram = \frac{Count(w_{i-1}\, w_i)}{Count(w_{i-1})}$$

$$Uni - Gram = \frac{Count(w_i)}{N}$$

Literature review

Aniket Dalal et al. (2006) [1], developed a Hindi Part-of-Speech Tagger using Maximum Entropy Approach. This paper presents a statistical Part of Speech (POS) tagger for a morphologically rich language: Hindi. This tagger employs the maximum entropy Markov model with a rich set of features capturing the lexical and morphological characteristics of the language. The feature set was arrived at after an exhaustive analysis of an annotated corpus. The morphological aspects were addressed by features based on information retrieved from a lexicon generated from the corpus, a dictionary of the Hindi language and a stemmer.

This system processes data in two phases. In the first phase, resources necessary for the tagging phase was generated. The generated resources include the list of unique words in the training corpus, called *lexicon*, and a restricted dictionary called *TinyDict*. A lexicon generator was run on the training corpus to create the lexicon. In the lexicon, along with the word, a flag was stored to indicate occurrence of the word as a proper noun in the training corpus. If a word appears with the tag proper noun at least once in the training corpus, then this flag is true. TinyDict stores information about the list of possible POS tags according to the dictionary. Another resource that is generated in pre-processing phase is the list of suffixes for all words using stemmer.

The system was evaluated over a corpus of 15,562 words developed at IIT Bombay with 27 different POS tags. They performed 4-fold cross validation on the data, and the system achieved the best accuracy of 94.89% and an average accuracy of 94.38%. This work shows that linguistic features play a critical role in overcoming the limitations of the baseline statistical model for morphologically rich languages.

The system used two measures to evaluate the performance of system, namely, per word tagging accuracy and sentence accuracy. Per word tagging accuracy is the ratio of number of words that are tagged correctly to the number of words present in the text. Sentence accuracy represents the percentage of sentences for which the tag sequence assigned by the system matches the true tag sequence. If all the words in a sentence are assigned correct tags, then the sentence is said to be correctly tagged. Sentence accuracy is the ratio of correctly tagged sentences to the number of sentences present in the text. The best context window was determined empirically. The context window consists of two words on either side of the current word and combination of these POS tags. The best per word tagging accuracy of the tagger for this context window without any other feature was 77.73%. The best result 85.59% was obtained with the context window consisting of the POS tags of the previous word, the current word and the next word.

Manish Shrivastava et al. (2008) [2] developed a Hindi Part of Speech Tagger using Hidden Markov Model. This paper presents a simple HMM based POS tagger, which employs a naive (longest suffix matching) stemmer as a pre-processor to achieve reasonably good accuracy of 93.12%. This method does not require any linguistic resource apart from a list of possible suffixes for the language. This list can be easily created using existing machine learning techniques.

The core idea of this approach is to "explode" the input in order to increase the length of the input and to reduce the number of unique types encountered during learning. This in turn increased the probability score of the correct choice while simultaneously decreasing the ambiguity of the choices at each stage. This also decreases data sparsity brought on by new morphological forms for known base words.

The corpus used for the training and testing purposes contains 66900 words. This data was 'exploded' resulting in a new corpus of 81751 tokens which was divided into 80% and 20% parts. The test set contained 13500 words which resulted in an exploded test set of 16000 tokens (stem and suffix tokens). The accuracy was calculated after imploding the output considering the assigned tag of the stem as the correct tag. This data was sourced from various domains including news, tourism and fiction.

Himanshu Agarwal et al. (2006) [3] developed a Part of Speech Tagging and Chunking using Conditional Random Fields. This paper present a CRF (Conditional Random Fields) based Part of Speech tagger for Hindi. Various experiments were carried out with various sets and combinations of features which mark a gradual increase in the performance of the system throughout building process. Apart from CRF based learning using the CRF package "CRF++, Yet Another CRF Package", a morph analyzer is used to provide extra information like root word and possible POS tags for training. With training on 21000 words with the best feature set, the CRF based POS tagger is 82.67% accurate.

For training the POS tagger we use the Hindi morph analyzer to get the root-word and possible pos tags for every word in the corpus. Along with the root-word and suggested pos tags other information like suffixes, word length indicator and presence of special characters is added to the training data. The data is then trained using "CRF++, Yet another CRF package" on a set detailed features and their combination. For running the system on unannotated input, a similar data has to be prepared before the system can predict.

Smriti Singh et al. (2006) [4] developed a Part of Speech Tagger using decision tree based learning. This paper presents a POS tagger for Hindi- the national language of India, spoken by 500 million people and ranking 4th in the world. This methodology uses locally annotated modestly sized corpora (15,562 words), exhaustive morphological analysis backed by high coverage lexicon and a decision tree based learning algorithm- CN2 (Clark and Niblett, 1989).

Morphology driven tagger makes use of the affix information stored in a word and assigns a POS tag using no contextual information. Though, it does take into account the previous and the next word in a VG to correctly identify the main verb and the auxiliaries, other POS categories are identified through lexicon lookup of the root form.

The tests were performed on contiguous partitions of the corpora (15,562 words) that are 75% training set and 25% testing set. The results are obtained by performing a 4-fold cross validation over the corpora. The average accuracy of the learning based (LB) tagger after 4-fold cross validation is 93.45%.

Smriti Singh in 2010 proposed a POS tagging methodology which can be used by languages having lack of resources [1]. The POS tagger was built based on hand - crafted morphology

rules and does not involve any sort of learning or disambiguation process. The system makes use of locally annotated modestly - sized corpora of 15,562 words, exhaustive morphological analysis backed by high - coverage lexicon and a decision tree based learning algorithm (CN2). The system uses Lexicon lookup for identifying the other POS categories. The performance of the system was evaluated by a 4-fold cross validation over the corpora and found 93.45% accuracy.

Nidhi Mishra et al. (2011) [5] developed a Part of Speech Tagger using Rule Based approach. This system was designed to be useful for the linguists, Hindi language learners. This system scans the Hindi corpus and then extracts the sentences and words from the given Hindi corpus. Finally Display the tag of each Hindi word like noun tag, adjective tag, number tag, verb tag etc. and search tag pattern from database

Methodology

1.Problem definition

After studying various research papers we have concluded that Part of Speech tagger is an important part of the Natural Language Processing applications. Accuracy of different Natural Language Processing applications depends upon the accuracy of Part of Speech Tagger. Most of the Part of Speech taggers have been developed using statistical approach for different Indian languages. Punjabi is one of the official language of India. Punjabi is the official

language of Indian states like Punjab, Haryana, and Delhi and well understood by many other northern Indian regions. Punjabi is also a trendy language in Pakistani Punjab region. In India, gurmukhi script is used to write Punjabi it means from Gurus mouth and in Pakistan Shahmukhi is used means from the kings mouth.

Objectives 1.To develop annotated POS data.

2. To develop an algorithim to extract rules from annotated text.

3. To develop rule based pos algorthim .

4. To evaluate the system

Following rules are used to tag words:

Bigram Rules	Probability Values
JJ QT_QTC	0.0171102661596958
QT_QTC N_NN	0.516320474777448
N_NN PSP	0.389545974818771
PSP N_NN	0.495580110497238
PSP QT_QTC	0.0613259668508287
QT_QTC QT_QTC	0.0979228486646884
QT_QTC V_VM_VF	0.00890207715133531
V_VM_VF N_NNP	0.0193370165745856
N_NNP RD_PUNC	0.132986627043091
RD_PUNC QT_QTC	0.0849420849420849
QT_QTC N_NNP	0.186943620178042
N_NNP RD_SYM	0.0616641901931649
RD_SYM N_NN	0.174242424242424

Results

We have constructed three test data sets for testing. These test data sets are composed from different websites of Punjabi. These data sets enclosed the news from different domains and also contain essay and small stories.

Test Cases

Test No.	Domain	No. of words
Test Case 1	News	6,123
Test Case 2	Essay	3,246
Test Case 3	Short Stories	982

Evaluation Metrics

The evaluation metrics for the data set is precision, recall and F-Measure. These are defined as following:-

$$Recall = \frac{Number\ of\ Correct\ answers\ Given\ by\ system}{Total\ Number\ of\ Words}$$

$$Precision = \frac{Number\ of\ Correct\ answers}{Total\ Number\ of\ Words}$$

$$F - Mesure = 2 * \frac{\mathrm{PR}}{(\mathrm{R} + \mathrm{P})}$$

Evaluation of Test Cases

Test Case No	Recall	Precision	F-Meaure
Test Case 1	0.931	0.909	0.92
Test Case 2	0.90	0.89	0.89
Test Case 3	0.922	0.87	0.89

CONCLUSION

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. Taggers may use several kinds of information like dictionaries ,rules

etc. pos tagging is very essential in natural language processing. The problem of tagging is widely studied in informal retrieval. supervised learning approach provides better results. Based on our training data we achieve desirable results. The other languages like English is a vast language as one can find as much resources as they want for training and testing. Based on our training data we are capable in achieving the satisfactory results.

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.