POS TAGGING OF PUNJABI LANGUAGE USING HIDDEN MARKOV MODEL

[1]Sapna Kanwar, [2]Mr Ravishankar, [3]Sanjeev Kumar Sharma
[1]LPU, Jalandhar, [2]Lecturer, LPU, Jalndhar, [3]Associate professor, B.I.S College of Engineering and Technology, Moga – 142001, India

Abstract : *POS tagger is the process of assigning a correct tag to each word of the sentence. We attempted to improve the accuracy of existing Punjabi POS tagger. This POS tagger lacks in resolving the ambiguity of compound and complex sentences. A Bi-gram Hidden Markov Model has been used to solve the part of speech tagging problem. An annotated corpus was used for training and estimating of HMM parameter. Maximum likelihood method has been used to estimate the parameter. This HMM approach has been implemented by using Viterby algorithm.*

1 INTRODUCTION

Part-of-speech (POS) Tagging is a process that attaches each word in a sentence with suitable tag from a given set of tags. It is one of the major activities performed in a typical natural language processing application such as speech recognition, information extraction, machine translation, grammar checking and word sense disambiguation etc. This paper explores part-of-speech tagging for the Punjabi language, a member of the Modern Indo-Aryan family of languages. There are two approaches of POS taggers: rule based and trained one. In the rule based approach, a knowledge base of rule is developed by linguistic to define precisely how and where to assign the various POS tags. This approach has already been used to develop the POS tagger for Punjabi language. In the trained approach, statistical language model are built, refined and used to POS tag the input text automatically. One of the robust approaches in statistical models is the use of Hidden Markov

Model (HMM). HMM is one of the distinguished probabilistic models used to work out a no of different problems and hence also repeatedly used in language processing problems. Especially for the case of disambiguation issues, HMM has been effectively utilized to find out most probable state sequence for a particular sentence. In this paper we report on the building and use of HMM based POS tagger for Punjabi. We have favored Hidden Markov Model over other statistical models for a no of reasons. First HMM models make use of History events in assigning the current event some probability value and that suits our approach philosophy. Second, HMM is superior to other models with regard to training speed. Hence HMM is suitable for applications that have to process large amounts of text. Recently a rule based POS tagger was developed that shows an accuracy of nearly 80%.

2 OVERVIEW OF PUNJABI LANGUAGE

Punjabi language is a member of the Indo-Aryan family of languages, also known as Indic languages. Other members of this family are Hindi, Bengali, Gujarati, and Marathi etc. Indo-Aryan languages form a subgroup of the Indo-Iranian group of languages, which in turn belongs to Indo-European family of languages. Punjabi is spoken in India, Pakistan, USA, Canada, England, and other countries with Punjabi immigrants. It is the official language of the state of Punjab in India. Punjabi is written in „Gurmukhi script in eastern Punjab (India), and in „Shahmukhi script in western Punjab (Pakistan).

3 PREVIOUS WORK

A rule based part-of-speech tagging approach was used for Punjabi, which is further used in grammar checking system for Punjabi [14]. This is the only tagger available for Punjabi Language. A part-of-speech tagging scheme based entirely on the grammatical categories taking part in various kinds of agreement in Punjabi sentences has been proposed and applied successfully for the grammar checking of Punjabi [14]. This tagger uses handwritten linguistic rules to disambiguate the part-of-speech information, which is possible for a given word, based on the context information. A tagset for use in this part-of-speech tagger has also been devised to incorporate all the grammatical properties that will be helpful in the later stages of grammar checking based on these tags. This part-of-speech tagger can be used for rapid development of annotated corpora for Punjabi. There are around 630 tags in this fine-grained tagset. This tagset includes all the tags for the various word classes, word specific tags, and tags for punctuations. During tagging process with proposed tagger, 503 tags out of proposed 630 tags were found in 8-million words corpus of Punjabi, which was collected from online sources. For disambiguation of POS tags rule-based approach was used. A database was designed to store the rules, which is used by rule based disambiguation approach. The texts with disambiguated POS tags are than passed for marking verbal operators. Four operator categories have been established to make the structure of verb phrase more understandable. During this step the verbal operators are marked based on their position in the verb phrase and the forms of their proceeding words. A separate database was maintained for marking verbal operator.

4 HIDDEN MARKOV MODEL

A hidden Markov model (HMM) is a statistical construct that can be used to solve classification problems that have an inherent state sequence representation. The model can be visualized as an interlocking set of *states.* These states are connected by a set of *transition probabilities,* which indicate the probability of traveling between two given states. A process begins in some state, then at discrete time intervals, the process "moves" to a new state as dictated by the transition probabilities. In an HMM, the exact sequence of states that the process generates is unknown (i.e., *hidden).* As the process enters each state, one of a set of *output symbols* is emitted by the process. Exactly which symbol is emitted is determined by a probability distribution that is specific to each state. The output of the HMM is a sequence of output symbols.

Basic Definitions and Notation

According to (Rabiner, 1989), there are five elements needed to define an HMM:

1. N, the number of distinct states in the model. For part-of-speech tagging, N is the num$_{ij}$ber of tags that can be used by the system. Each possible tag for the system corresponds to one state of the HMM.
2. M, the number of distinct output symbols in the alphabet of the HMM. For part-ofspeech tagging, M is the number of words in the lexicon of the system.
3. A = $\{a_{ij}\}$, the state transition probability distribution. The probability *a* is the probability that the process will move from state i to state j in one transition.

For part-of-speech tagging, the states represent the tags, so a_{ij} is the probability that the model will move from tag t_i to t_j — in other words, the probability that tag t_j follows t_i. This probability can be estimated using data from a training corpus.

4. *B = {b_j(k)),* the observation symbol probability distribution. The probability *b_j(k)* is the probability that the k-th output symbol will be emitted when the model is in state j. For part-of-speech tagging, this is the probability that the word W_k will be emitted when the system is at tag t_j (i.e.,*$P(W_k|t_j)$).* This probability can be estimated using data from a training corpus.

5. “ = {“$_i$}, the initial state distribution. “$_i$ is the probability that the model will start in state i. For part-of-speech tagging, this is the probability that the sentence will begin with tag ti. When using an HMM to perform part-of speech tagging, the goal is to determine the most likely sequence of tags (states) that generates the words in the sentence (sequence of output symbols). In other words, given a sentence V, calculate the sequence U of tags that maximizes *P(V/U).* The Viterbi algorithm is a common method for calculating the most likely tag sequence when using an HMM.

The presented model is a type of first order HMM, also referred to as bigram POS tagging. For POS-tagging problem presented Hidden Markov Model is composed of two probabilities: lexical (emission) probability and contextual (transition) probability (Samuelsson, 1996).

$$(t_1,....,1_n)^0 = \underset{t_1....t_n}{\arg\max}\, P(t_1,....,1_n) \,|\, (W_0,....,W_n)$$

Using Baye’s law above equation can be rewritten as:

$$P(t_1,.....,t_n \,|\, W_1,.....,W_n) = P(t_1,.....,t_n)x\frac{P(W_1,.....,W_n \,|\, t_1,.....,t_n)}{P(W_1,.....,W_n)}$$

$$(t_1,.....,t_n)^0 = \underset{t_1,.....,t_n}{\arg\max}\, P(t_1,.....,t_n)xP(W_1,.....,W_n \,|\, t_1,.....,t_n)$$

$$(t_1,.....,t_n)^0 = \underset{t_1,.....,t_n}{\arg\max}\, P(t_1,.....,t_n)xP(W_1,.....,W_n \,|\, t_1,.....,t_n)$$

$$= \underset{t_1,....t_n}{\text{argmax}} \prod_{i=1}^{=} (\underbrace{P(t_i|t_{i-1})}_{\text{Transition probability}} * \underbrace{P(w_i|t_i)}_{\text{Emission probability}})$$

5 Viterbi Algorithm

We now know how to derive the probabilities needed for the Markov model, and how to calculate P(T | W) for any particular (T, W) pair. But what we really need is to be able to find the most likely T for a particular W. 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. The pseudo code of the algorithm is shown bellow.

```
For i = 1 to Number_of_Words_in_Sentence
    for each state c “ Tag_Set
        for each state b “ Tag_Set
            for each state a “ Tag_Set
                For the best state sequence ending in state a at time (i -2), b at
                time (i-1), compute the probability of that state sequence going to
                state c at time i.
            end
        end
    end
```

Determine the most-probable state sequence ending in state c at time i end So if every word can have S possible tags, then the Viterbi algorithm runs in O(S3*|W|) time, or linear time with respect to the length of the sentence.

6 MAXIMUM LIKELIHOOD ESTIMATION: Maximum likelihood is one of the simplest ways to compute probabilities through relative frequencies. In case of HMM, we estimate probability distribution variables for the model parameters ë = (A, B, ð) in the training corpus (Blunsom, 2004; Padró and Padro, 2004), as given below.

$$\pi_i = \frac{C(q_1 = t_i)}{C(q_1)}$$

$$a_{ij} = \frac{C(t_i, t_j)}{C(t_i)}$$

$$b_j(k) = \frac{C(W_k, t_j)}{C(t_j)}$$

Where:

$C(t_i, t_j)$ = Denotes the count (or number of times) that state t_j is followed by state t_i

$C(w_k, t_j)$ = Denotes the number of times w_k tagged with t_j

7 EXPERIMENTAL EVALUATION

The accuracy of any Part of Speech tagger is measured in terms of the accuracy i.e. the percentage of words which are accurately tagged by the tagger. This is defined as belows:

Accuracy = Total no of words having correct tag / total no of words tagged

For evaluation of the proposed tagger, a corpus having texts from different genres were used. The outcome was manually evaluated to mark the correct and incorrect tag assignments. 20,000 words collected randomly from a 4 million corpus of Punjabi were manually evaluated and are grouped into two genres.

Test set	Size (No of words)	Accuracy
A	10000	84.9%A
B	10000	87.6%

8 CONCLUSIONS AND FUTURE WORK

In this study, we proposed the initial implementation of HMM to one of the partially free word and morphology rich language Punjabi. During experimental results we note that the general HMM based method doesn't perform well due to data sparseness problem. In future, we intend to develop novel methods to improve overall accuracy and specifically unknown words in Punjabi and other word-free languages. We aim to find out ways to improve the language model behavior without increasing the training corpus and by integrating linguistics knowledge.

References

[1] Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke, "Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach", In Proceeding of the NLPAI Machine Learning Competition, 2006.

[2] Antony P.J, Santhanu P Mohan, Soman K.P,"SVM Based Part of Speech Tagger for Malayalam", IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 339-341, 2010

[3] Agarwal Himashu, Amni Anirudh," Part of Speech Tagging and Chunking with Conditional Random Fields" in the proceedings of NLPAI Contest, 2006

[4] Bird, S., E. Klein and E. Loper, 2007. Natural language processing in python. University of Pennsylvania, Nltk. Sourceforge. http://mail.python.org/pipermail/python-list/2007-May/442489.html.

[5] Blunsom, P., 2004. Hidden Markov models. Technical Report.

[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] Ekbal, Asif, Haque, R. and S. Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field Approach", In Proc. of 3rd IJCNLP, 51-55, 2008

[11] Ekbal, A. Bandyopadhyay, S., "Part of Speech Tagging in Bengali Using Support Vector Machine", ICIT- 08, IEEE International Conference on Information Technology, pp. 106-111, 2008

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

[13] 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

[14] Gurpreet Singh, "Development of Punjabi Grammar Checker, Phd. Dissertation, 2008

[15] Jurafsky D and Marting J H, Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Pearson Education Series 2002

[16] James Allen, Natural Language Understanding, Benjamin/ Cummings Publishing Company, 1995

[17] Jes´us Gim´enez and Llu´ýs M'arquez., SVMTtool:Technical manual v1.3, August 2006

[18] John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings *of the 18th International Conf. on Machine Learning*, pages 282–289.Morgan Kaufmann, San Francisco, CA.

[19] Kudo, T and Matsumoto, "Chunking with Support Vector Machines", In Proc. of NAACL, 192-199, 2001.

[20] Lafferty, J., McCallum, A., and Pereira, F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In Proc. of the 18th ICML 01, 282- 289, 2001.

[21] Linda Van Guilder (1995) Automated Part of Speech Tagging: A Brief Overview Handout for LING361, Fall 1995 Georgetown University

[22] Manju K., Soumya S., Sumam Mary Idicula, "Development of a POS Tagger for Malayalam - An Experience," artcom, pp.709-713, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009

[23] Manish Shrivastava and Pushpak Bhattacharyya, Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge, International Conference on NLP (ICON08), Pune, India, December, 2008 Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008

[24] PVS Avinesh, G Karthik, "Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning" in the proceedings of NLPAI Contest, 2006

[25] Ratnaparkhi, A., “A Maximum Entropy Part of Speech Tagger”, In Proc. of the EMNLP Conference, 133-142, 1996

[26] RamaSree, R.J, Kusuma Kumari, P., “Combining Pos Taggers For Improved Accuracy To Create Telugu Annotated Texts For Information Retrieval”, 2007, Available at http://www.ulib.org/conference/2007/RamaSree.pdf

[27] Sumam Mary Idicula and Peter S David, A Morphological processor for Malayalam Language, South Asia Research, SAGE Publications, 2007

[28] Sandipan Dandapat, Sudeshna Sarkar, Anupam Basu,” Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario”, Proceedings of the Association for Computational Linguistic, pp 221-224, 2007

[29] Scott, M.T. and M.P. Harper, 1999. A second-order Hidden Markov Model for part-of-speech tagging. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Jun. 20-26, College Park, Maryland, pp: 175-182.

[30] Sigletos, G., G. Paliouras and V. Karkaletsis, 2002. Role identification from free text using hidden Markov models. Proceedings of the 2nd Hellenic Conference on AI: Methods and Applications of Artificial Intelligence, 2002, Springer Verlag, pp: 167-178.

[31] S. Singh , K. Gupta , M. Shrivastava and P. Bhattacharya, “Morphological Richness Offsets Resource Demand-Experiences in Constructing a POS Tagger for Hindi”, In Proc. of COLING/ACL, 779-786, 2006

[32] Singh Mandeep, Lehal Gurpreet, and Sharma Shiv, 2008. “A Part-of-Speech Tagset for Grammar Checking of Punjabi”, published in The Linguistic Journal, Vol 4, Issue 1, pp 6-22

[33] Smriti Singh, et.al,” Morphological Richness Offsets Resource Demand-Experiences in Constructing a POS Tagger for Hindi”, in the proceedings of COLING/ACL, pp. 779-786, 2006