International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 13 • Issue 1 pp. 12-15 Jan - June 2021 www.csjournals.com

Development of POS Tag Set for the Dogri Language using SMT

Shubhnandan S. Jamwal
PG Department of Computer Science and IT, University of Jammu, Jammu

Abstract: Natural Language processing which is a part of AI is mainly concerned with the development of computational models and tools for processing the task of the NLP. The development of the Part of Speech Tagging remains a challenge for the low resourced language like Dogri. But the problem of the POS tag development is well studied topic and also one of the most fundamental preprocessing steps for any language processing in NLP. POS tagging of Dogri language is a necessary component for the development of any language. In this paper, the SMT approach based on HMM is studied and implemented for the development of the Dogri POS tagged corpus. It is observed that the context of the training and test data is important to achieve the desired goal. The level of accuracy remains at only 83% when the context is same and reduced to 59% when the context is changed.
Key words: Dogri, POS, SMT, HMM, Tagged Corpus.

Introduction

Dogri language is spoken by people residing in Jammu and Kashmir, and in some parts of Punjab and Himachal Pradesh. Dogri is one of the 22 languages scheduled in constitution of India and now recognized and official language of the J&K UT. The modern era of the digital world, the people of the India want to communicate in the regional languages on social media platforms. Therefore, the development of the regional languages and the NLP tasks is very important for the regional languages.

The present NLP system in the modern digital era plays very important role every individual wants the information on click preferably in the local language. As per the desire of the netizens the information must be retrieved by search engines on the click. Web search engines are developing different techniques to make searching faster and retrieves information efficiently. The dream of artificial intelligence research in the language pair translation begins from the automatically translating documents between two languages. This remains as one of the oldest pursuits of natural language processing. Since very powerful computing facility is available now a day's, NLP and AI are going side by side in pace. Neural machines and Statistical Machine Translation based on the corpus approaches requires a parallel corpus to learn a model [1][2] and are very much in progress of development in Indian languages pairs. For the development of the machine based on corpus approach we need to have parallel corpora which are different from normal text corpora. The text in parallel corpora is aligned line by line so that accuracy can be obtained. But the languages which have rich morphologically and are inflectional makes the NLP application task tough.

Literature Review

K. M. Shivakumar, N. Shivaraju, V. Sreekanta and D. Gupta [3] presented the comparison between Statistical Machine Translation (SMT) model with linguistic and non-linguistic data models for English to Kannada languages. The experiments shows an improvement in Bleu-Score for Factored MT system against Baseline MT system for English to Kannada SMT. Kannada fonts can take ten different forms in representing a word any change of a font variant in word leads to change in meaning of the word. We model these morphological variants of Kannada lemma words, their variants and PoS as Factors in our MT System. Z. Yang, M. Li, Z. Zhu, L. Chen, L. Wei and S. Wang [4] proposed a method that adopts morphological information as the features of the maximum entropy based phrase reordering model for Mongolian-Chinese SMT. By taking advantage of the Mongolian morphological information, they add Mongolian stem and affix as phrase boundary information and use a maximum entropy model to predict reordering of neighbor blocks. To some extent, our method can alleviate the influence of reordering caused by the data sparseness. In addition, they further add

part-of-speech (POS) as the features in the reordering model. Experiments show that the approach outperforms the maximum entropy model using only boundary words information and provides a maximum improvement of 0.8 BLEU score increment over baseline. K. Phodong and R. Kongkachandra[5] presented a method to improve Thai-English word alignment in statistical machine translation (SMT) for interrogative sentences in a parallel corpus. They utilize the Thai and English grammatical knowledge i.e. tense, part of speech (POS), and question inversion pattern. The proposed method handles the difference of Thai and English interrogative sentences using sentence transformation, interrogative grammatical attribute extraction, and interrogative grammatical attribute annotation. This method works as a pre-processing of GIZA, a standard word co-occurrence alignment tool in SMT. They hypothesize that using grammatical knowledge as a pre-processing of GIZA can provide higher accuracy. The data set for the experiment is composed of 43,500 interrogative sentences to compare alignment result between interrogative sentences attached an interrogative grammatical label and interrogative sentences unattached interrogative grammatical label. The experimental results yield 95% of accuracy with significant improvement than the conventional one. With the increasing accuracy of word alignment, the translation accuracy is consequently improved. S. D. Makhija[6] discussed various stemmers which are developed for different languages and aims to develop a stemmer for Devanagari script based Sindhi language which strips the prefixes and suffixes from inflected word to their stem or root.

S. K. Nambiar, A. Leons, S. Jose and Arunsree [7] described the methods to build a Part of speech tagger by using hidden markov model. Supervised learning approach is implemented in which, already tagged sentences in malayalam is used to build hidden markov model. J. Srivastava and S. Sanyal [8] described the use of part of speech (POS) tag to improve the performance of statistical word alignment. The approach proposed works well with small size of the corpus. Experiments were conducted on TDIL sample tourism corpus of 1000 sentences for English-Hindi language pair. Out of these 1000 sentences 950 sentences are used for training and 50 sentences are used for testing. F-measure is increased by approximately 4% and Alignment Error Rate (AER) decreased by approximately 4% in comparison to baseline system for word alignment GIZA++. R. R. Deka, S. Kalita, K. Kashyap, M. P. Bhuyan and S. K. Sarma [9] presented the performances of two existing tagging techniques for Assamese language that is, Conditional Random Field and Trigrams'n Tag to study the efficiency of the models in tagging of Parts-of-speeches for the language and the study might also help the Natural Language Processing (NLP) researchers in getting an understanding and analyzing the performance standard and effectiveness of this two existent models for POS tagging task for other morphologically rich Indian languages. A. Paul, B. S. Purkayastha and S. Sarkar [10] used Hidden Markov Model (HMM) based Part of Speech (POS) tagging for Nepali language. HMM is the most popular used statistical model for POS tagging that uses little amount of knowledge about the language, apart from contextual information of the language. The evaluation of the tagger has been done using the corpora, which are collected from TDIL (Technology Development for Indian Languages) and the BIS tagset of 42 tags. Tagset has been designed to meet the morph-syntactic requirements of the Nepali language. Apart from corpora and the tagset, python programming language and the NLTK's (Natural Language Toolkit) library has been used for implementation. The tagger achieves accuracy over 96% for known words but for unknown words, the research is still continuing. The previous work reported in the Dogri language is for the generation of the verbs of the language [11] and the automatic generation of the Noun using Machine learning [12]. English-Dogri parallel corpus has also been prepared for machine translation from one language into another language and Moses, which is a statistical machine translation system has been used to train translation models for any language pair[13]. A rule based transliteration system has also been developed for detection of proper nouns for Dogri to English[14] and stemmer[15] for the Dogri language is also developed.

SMT-POS for Dogri

In the modern world large volume of NLP works are going on. Different models and algorithms are also used for developing the POS tags. The tags are used for each word having different parts of speech. In this paper we are using the BIS tags for the development of the tagged corpus. Part-of-speech tagging is the process of assigning a part-of-speech to each word in part-of-

speech tagging a text. The tags are applied to the new corpus word based on the statistical information of the word in the existing corpus as shown the fig 1.

AhlaMá	1	
KadZerI	6	
saMjIxA	5	
KalakawE		1
brija	4	
pIE	14	
baMXana	3	
tEMte	2	
KubBI	3	
ke	26	
vanAspawI		2
bETA	19	
CAwiyA	2	
hAusiMga		1
caMxe	1	
karoAMxiyAM		3
mUMhE	1	
mAlakana		3
pawIlA	1	
liKiyAM	5	
soAla	78	
jAo	32	
poWI	1	
xaswaKawa		3
ciwE	2	
jAhgArYarYa		1
nohAra	6	
turE	1	
KaDAMxA	1	
cuppa-cApa		11
bajaxe	2	
KeDaxA	3	

Fig 1: No of the occurrences of a particular word in the corpus

In this framework one can develop the XML based POS tags also. The input to the engine is the tag and the sequence W1, W2,..., Wn of words and a BIS tag set, and the output is a sequence words (W1, W2,..., Wn) with tags like राम/NP जागतें/NC कन्ने/PP पानी/NN पीन/VB जंदा/VM ऐ/VA.

Results

In this paper, we have used the Hidden Markov Model for applying the POS tags. HMM is a class of probabilistic graphical model which is selected for the testing and applying tags on the text of Dogri data. It is predicting the sequence of tags from the observed variables. In our experiment the text is composed of the 1150 words and the tagged corpus on which the HMM is based is composed of the ten thousand words. The level of accuracy remains at only 83% when the context of the training and text data remains same and when the context is change the level drops to 59%. This is perhaps because of the highest level of the inflections in the language Dogri.

Conclusion

POS, Named entities, Stemmers, Chunkers, etc. remain useful tools for the development of the NLP application. The building of the NLP applications starts from the processing of the sentence structure and its meaning. The HMM remains the best choice for the development of the POS tagged data based on the SMT. But it is observed that if the context of the test data is changed the level of the accuracy is dropped to a great extent and if the HMM is to be adopted for then the volume of the training data needs to be increased based on the different contexts. The researchers are also proposing the Viterbi algorithm which is based on the dynamic programming for finding the most likely sequence of hidden states.

References

[1] Koehn P. Statistical Machine Translation. New York: Cambridge University Press. 2010.
[2] Peng L. A Survey of Machine Translation Methods. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2013; 11(12): 7125-7130.
[3] K. M. Shivakumar, N. Shivaraju, V. Sreekanta and D. Gupta, "Comparative study of factored SMT with baseline SMT for English to Kannada," 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, pp. 1-6, doi: 10.1109/INVENTIVE.2016.7823217.
[4] Z. Yang, M. Li, Z. Zhu, L. Chen, L. Wei and S. Wang, "A maximum entropy based reordering model for Mongolian-Chinese SMT with morphological information," 2014 International Conference on Asian Language Processing (IALP), 2014, pp. 175-178, doi: 10.1109/IALP.2014.6973484.
[5] K. Phodong and R. Kongkachandra, "Improving Thai-English word alignment for interrogative sentences in SMT by grammatical knowledge," 2017 9th International Conference on Knowledge and Smart Technology (KST), 2017, pp. 226-231, doi: 10.1109/KST.2017.7886115.
[6] S. D. Makhija, "A study of different stemmer for Sindhi language based on Devanagari script," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 2326-2329.
[7] S. K. Nambiar, A. Leons, S. Jose and Arunsree, "Natural Language Processing Based Part of Speech Tagger using Hidden Markov Model," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2019, pp. 782-785, doi: 10.1109/I-SMAC47947.2019.9032593.
[8] J. Srivastava and S. Sanyal, "POS-based word alignment for small corpus," 2015 International Conference on Asian Language Processing (IALP), 2015, pp. 37-40, doi: 10.1109/IALP.2015.7451526.
[9] R. R. Deka, S. Kalita, K. Kashyap, M. P. Bhuyan and S. K. Sarma, "A Study of T'nT and CRF Based Approach for POS Tagging in Assamese Language," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 600-604, doi: 10.1109/ICISS49785.2020.9315939.
[10] A. Paul, B. S. Purkayastha and S. Sarkar, "Hidden Markov Model based Part of Speech Tagging for Nepali language," 2015 International Symposium on Advanced Computing and Communication (ISACC), 2015, pp. 149-156, doi: 10.1109/ISACC.2015.7377332.
[11] Jamwal S.S., Gupta P., Sen V.S. (2021) Hybrid Model for Generation of Verbs of Dogri Language. In: Singh T.P., Tomar R., Choudhury T., Perumal T., Mahdi H.F. (eds) Data Driven Approach Towards Disruptive Technologies. Studies in Autonomic, Data-driven and Industrial Computing. Springer, Singapore. https://doi.org/10.1007/978-981-15-9873-9_39.
[12] Shubhnandan S. Jamwal, Named Entity Recognition for Dogri using ML, International Journal of IT & Knowledge Management, Jan-June 2017 Volume-10, Number-2 pp. 141-144.
[13] Singh, Avinash & Kour, Asmeet & Jamwal, Shubhnandan. (2016). English-Dogri Translation System using MOSES. Circulation in Computer Science. 1. 45-49. 10.22632/ccs-2016-251-25.
[14] Tejpaul S. Sasan, Dr. Shubhnandan S. Jamwal, Transliteration of Name Entities Using Rule Based Approach, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 6, Issue 6, June 2016.
[15] Gupta P., Jamwal S.S. (2021) Designing and Development of Stemmer of Dogri Using Unsupervised Learning. In: Marriwala N., Tripathi C.C., Jain S., Mathapathi S. (eds) Soft Computing for Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-16-1048-6_11