International Journal of Advances in Applied Sciences (IJAAS)
Vol. 8, No. 4, December 2019, pp. 264~268
ISSN: 2252-8814, DOI: 10.11591/ijaas.v8.i4.pp264-268

Survey of part-of-speech tagger for mixed-code Indian and foreign language used in social media

Bhushan Nikam
Department of Computer Science, Dr. D. Y. Patil ACS College, India

Article Info

Article history:

Received Apr 29, 2019
Revised Aug 28, 2019
Accepted Oct 6, 2019

Keywords:

Information Extraction
Machine Translation
POS tools

ABSTRACT

A Part-Of-Speech Tagger (POS Tagger) is a tool that scans the text in specific language and allocates chunks of speech to individual word (and another token), such as verb, adjective, nown etc., as more fine-grained POS tags are used in computational applications like 'noun-plural'. Basically, the goal of a POS tagger is to allocate linguistic (mostly grammatical) information to sub-sentential units, called tokens as well as to words and symbols (e.g. punctuation). This paper presents a survey of POS Tagger used for code-Mixed Indian and Foreign languages. Various methods, procedures, and features required to device POS Tagger for code-mixed foreign languages especially for Indian are studied and observations related to it are reported.

Corresponding Author:

Bhushan Nikam,
Department of Computer Science,
Dr. D. Y. Patil ACS College,
Sant Tukaram Nagar, Pimpri Colony, Pune, Maharashtra 411018, India.
Email: bhushannikam1973@gmail.com

1. INTRODUCTION

Community language of communication in social media is often combined in nature, where individuals counterfeit their regional dialectal with English and this technique is found to be extremely popular. Natural language processing (NLP) work towards to gather the data from these texts somewhere Part-of-Speech (POS) tagging performs a key title role in receiving the prosody of the inscribed text. One purpose of POS labeling is to disambiguate homonyms. Several kinds of information including dictionaries, lexicons, rules etc. use by taggers. Word may be a member of more than one category. Lexicons have type or types of a specific word. For example, a word address is both verb and noun. Taggers utilizes the probabilistic evidence to solve this indistinctness of actual word. As a preprocessor in text processing POS tagger can be used. Text retrieval and indexing requires POS information. Language processing needs POS tags to choose the pronunciation. For making tagged corpora POS tagger is also used.

Dialectal processing methods to code switched text was first accomplished in the early 1980s [1], whereas in social media text code-switching begun to be considered in the late 1990s [2]. Still, conventional texts code change was rare as to encourage ample curiosity by the computational dialectal research people, and it was first lately that, it emerges a study topic in its own right, with a code-switching workshop at EMNLP 2014 [3]. Solorio with Liu [4], projected a simple but well-designed solution of labeling mixed-code English-Spanish transcript twice - on one occasion for each language, a tagger - and then joining the outcome of the language-explicit taggers to get the optimal word-level tags [5]. For English-Hindi Mixed-Code Social Media Content, a POS Labeling System has been presented in [5]. Efforts has been performed on English-Bengal and English-Hindi data. Nelakuditi [6], performed, two different kinds of experiments, First, POS taggers based on machine learning and second is uniting POS taggers of individual languages [7].

POS tagger tool has been designed for various languages, but for code-mixed Indian and foreign Languages, very little work yet is performed with undesirable accuracy. This paper presents review of such work which is prepared into next four Sections. Section 2 and 5 specifies techniques used and approaches involved in the implementation of POS tagger for code-mixed Indian and foreign dialects. Section 3 summarizes efforts made to implement CM POS tagger for Indian Languages. Challenges to implement code-mixed POS tagger is presented in section 4.

2. VARIOUS APPROACHES AND TECHNIQUES USED TO IMPLEMENT CODE-MIXED POS TAGGER FOR INDIAN AND FOREIGN DIALECTS

India is homegrown to number of dialects. Language changes and variety in dialect prompt frequent mixing of code in India. Hence, Indians are polyglot by habituation with necessity, and frequently change mix tongues in social media circumstances, that possess additional problems for automatic Indian social media text processing. Requirement for any kind of NLP applications especially in this context Code-Mixed Part-of-speech (CM-POS) labelling is essential. Relating to it, I present a report on various POS tagger approaches and techniques used to implement code-mixed POS tagger for Indian and foreign Languages.

Jamatia and Das [5] experimented by using classification algorithms based on four machine learning technique to the undertaking exercise: Conditional Random Fields (CRF), with Sequential Minimal Optimization (SMO), Naïve Bayes (NB), and Random Forests (RF). For the Conditional Random Fields they tried the MIRALIUM[1] application, whereas the other three were the applications in WEKA[2] and reported effectuation on the complete dataset (2,583 utterances), after 5-fold cross-validation of all the ML methods using both fine-grained (FG) and coarse-grained (CG) tag sets and noticed that all the ML methods have further problems with HI-EN alternation.

In the Machine learning based POS taggers experiment Nelakuditi et. al [6] used three types of Machine Learning techniques for designing the POS tagger viz, Support Vector Machines (SVM), Bayes classification (Bay) and Conditional Random Fields (CRF), with different groupings and distinctions. In second experiment of joining POS taggers of individual languages, CMU's Twitter POS tagger for English with POS tagger developed at LTRC, that is a part of the shallow parser tool[3] for Telugu were used and then finally reported accuracies.

Kamal Sarkar [7], developed HMM-based POS tagging system which is founded on Trigram Hidden Markov Model that uses data from the vocabulary, and some other word level attributes to improve the comment possibilities of the known along with unknown tokens. He gives in to scores for Hindi-English, Bengali-English and Tamil-English Language duos. His scheme has been skilled and tried on the datasets provided for ICON 2015 shared task. In the constrained mode, his technique gains average overall accuracy (averaged over all three language pairs) of 75.60% which is very close to other participating two systems (76.79% for IIITH and 75.79% for AMRITA_CEN) which ordered larger than his system. In the unrestricted mode, his system gets typical overall accuracy of 70.65% which is also nearby to the system (72.85% for AMRITA_CEN) that obtained average overall accuracy highest.

Vyas et. al [8] conducted three different experiments: In the first experiment, by assuming the language identities and normalized/transliterated forms of the words, POS tagging is performed. It gives an idea of the accuracy of POS tagging task, if normalization, transliteration and language identification could be done perfectly. Experiments have been conducted with two different POS taggers for English: the Stanford POS tagger and the Twitter POS tagger. In the next experiment, by assuming that only the language identity of the words are known for Hindi their own model is applied to generate the back transliterations. For English, Twitter POS tagger is applied directly to handle social media text. In the third experiment by assuming nothing is known, language identifier process is first applied, and based on the language detected, Hi transliteration module, and Hi POS tagger, or the English tagger is applied and also stated that though the matrix information is not used in any of their experiments, it could be potentially useful for POS tagging which could be explored in future.

For constrained and unconstrained training and result submission, Pimpale and Patel [9], used Stanford POS tagger and machine learning algorithm viz., Decision Tree J48, Decision Tree Random Forest, Naive Bayes and Multilayer Perceptron resp. By concluding, the method used is reporting well for constrained submission, but deficiency of the superiority working information doesn't allow doing ample with it, if they, use the distributed vector illustration of words in feature engineering, that allow them to use non-labeled data for working out.

As stated by Sequiera et. al [10], explored machine learning approaches for Hindi (Hi)-English (En) CM typescript from social media POS tagging starting with repetition of the trials specified in [8] along with [4], and reconfirming results on dataset. Extending the attributes set applied by Solorio and Liu [4] and doing numerous feature selection experiments, they proposed and conducted a POS-tagging and joint

language labeling task. Their observations show that, when there is a marginal upgrading due to use of some supplementary features, joint modeling pointedly damages the results.

Kamal Sarkar [11], also proposed a POS tagging system for social media texts. It is developed based on Conditional Random Fields (CRF) trained using a rich feature set that includes contextual features, orthographic features, punctuation features and word length features. He concluded that his system performs well across all three languages Bengali-English-Hindi pairs. He hoped that the proper choice of features along with the suitable grouping of machine learning algorithms would improve the performance of his system.

According to Sharma and Motlani [12], experimented code-mixed POS tagging of Indian social media text using machine learning techniques. Building a POS tagger using constrained system, give them an accuracy of 75.04%, after being estimated on the new test dataset. While by using other resources, namely an unconstrained system, POS tagger did better than the constrained system and gives 80.68% of accuracy. For training and testing of both type of systems they used ten-fold cross-validation method and computed the best model attribute values by undertaking a grid search over all the parameters of the attributes. Finally, for the other two pairs, namely BN-EN (Bengali-English) and TA-EN (Tamil-English), accuracy measured was 79.84% and 75.48% respectively using developed and submitted constrained systems. Pipeline approach, for language identification, Back-transliteration and POS tagging Sisodiya [13] respectively used, logistic based classifier and CRF, Google API, and CRF++ based Hindi POS tagger developed by IIT Kharagpur.

Singh and Kanskar [14] employed, controlled word-level classification with and without contextual signs, and sequence labeling using Conditional Random Fields, for implementation of a simple unconfirmed dictionary-based method. A modest dialectal discovery-based investigative used in which first, the text can be separated into portions of tokens belonging to a language, and then each portion be categorized according to its language and further labeled by the POS tagger for that dialectal. Linguistic finding and transliteration text is labeled through an English monolingual tagger and then selecting one out of two labels for a conversation based on some heuristics that was detected by several language detection techniques.

As stated by Ghosh et. al [15], they listed various steps involved in POS labeling task using CRF++ toolkit and Stanford POS Tagger, including chunking, lexicons for dominant languages. They also concluded that Bengali-English and Hindi-English results are more than that of Tamil-English because of difference in labels used in Tamil-English gold standard files.

Barman [16], divided the experiment into four parts viz., implementing, baselines for POS tagging, pipeline systems, their stacking systems and joint model. By performing with the data, five-fold cross-validation and reported normal cross-validation exactness with investigating the use of hand-crafted features and attributes that can be gained from monolingual POS taggers (stacking), performed researches with different groupings of these attribute sets. They described a trilingual code-mixed corpus with POS comment. Using state-of-the-art methods performing POS tagging and investigating the usage of factorial CRF (FCRF)-based joint model found that the best stacking method (S2) that practices the joint features, achieves better than the combine version (FCRF) and the systems with pipeline. They observed that combined modeling outperforms the systems with pipeline in their experimentations. FCRF fall late the best POS labeling system S2. Possibly, to achieve better performance than S2 more training data would help FCRF.

According to Gupta et. al [17], they proposed a system that practices a comprehensive set of features for POS labeling. The feature set was used to design a POS model. Conditional random field (CRF) is applied as the underlying classifier. CRF++, an employment of CRF is used to accomplish the experiment. As CRF++ uses a stated feature template, therefore to discover the optimal feature template a series of experiments were made on the training data set in a cross-validated way. However, they tune the feature pattern on English-Hindi data set only and used the optimal model for all these CM languages (English-Hindi, English-Bengali, and English-Telugu) pairs. Bhargava *et. al* [18, 19], experimented similar kinds of approaches to implement POS tagger for English-Telugu, English-Hindi, English-Bengali language pairs with a slight variation to achieve accuracies. Table 1 shows the summarizing efforts made to implement CM POS tagger for indian languages.

Table 1. Summarizing efforts made to implement CM POS tagger for indian languages

Languages	Year	Approches/ Algorithm	F1/ Accuracy
English-Hindi	2014	CRF++, Twitter POS Tagger	74.87%
English-Hindi	2015	CRF, SMO, NB, RF	64.91%
English-Telugu	2015	SVM, Bayes classification, CRF, CMU's Twitter POS tagger, POS tagger developed at LTRC	52.37%
Telugu Hindi, Bengal mixed with English	2016	Stanford POS tagger, Decision Tree J48, Decision Tree, RF, NB, Multilayer Perceptron	Ta+en 71.04, 48.03 Bn+En 75.46, Not Submitted Hi+En 71.11, 6.84
Hindi-English	2015	ML Algo. with several features set experiments, joint modeling	77.33%
Hindi-English, Bengali-English and Tamil-English	2015	HMM-based POS tagging method	In Constrained mode 75.60%. In an Unconstrained mode 70.65%
Bengali-English, Hindi-English and Telegu-English Bengali-English, Hindi-English and Telegu-English Bengali-English, Hindi-English and Telegu-English Bengali-English, Hindi-English and Telegu-English	2016	CRF trained using a rich feature set	79.99%
Hindi-English Bengali-English, Tamil-English	2015	ML approach, POS tagger using unconstrained system and constrained system	Unconstrained Hi-En system 80.68% & constrained system for other 77.60%
English-Hindi	2015	logistic based classifier and CRF, Google API, CRF++	classifier using the CRF model 84.48%.
Hindi-English	2016	the dictionary-based approach, CRF, monolingual tagger, language detection techniques	-
Bengali-English, Hindi-English, Tamil-English	2016	Stanford POS Tagger and CRF++ toolkit	75.22% accuracy in Bengali-English
English-Bengali-Hindi	2016	baselines for POS tagging, pipeline systems, stacking systems, factorial CRF based joint model	84.58% on monolingual and 81.78% in code-mixed sentences
English-Hindi, English-Bengali and English-Telugu	2016	Rule-based tagging, CRF, CRF++	--
English-Telugu, English-Hindi, English-Bengali	2016	RF and Extremely Randomized Tree	78.744 % in fined grained system 77.944 % for coarse-grained model
English-Telugu, English-Hindi, English-Bengali	2016	Ran-dom forest, Logistic Regression, and Nave Bayes Ran-dom forest, Logistic Regression, and Nave Bayes RF, Logistic Regression, NB	F-Measure of Coarse-Grained Data Set C U Telugu-English 80.06 77.7 Hindi-English 71.03 71.655 Bengali-English 71.03 71.83 coarse-grained tag sets with an accuracy of 80.6% Coarse-grained 80.6%

3. VARIOUS APPROACHES AND TECHNIQUES USED TO IMPLEMENT CODE-MIXED POS TAGGER FOR FOREIGN LANGUAGES

Efforts are not much more still be seen to implement code-mixed POS tagger for foreign languages. Solorio and Liu [4] just predicted potential code alternation points, in the growth of extra accurate systems for processing code-mixed English-Spanish language. Such mixing of languages is rarely found all over the world, other than in India.

4. CHALLENGES TO IMPLEMENT CODE-MIXED POS TAGGER

Building Code-Mixed POS (CM-Part of Speech) taggers for Indian dialects is a particularly interesting problem in computational linguistics due to a lack of accurately glossed training corpora. More cultured language processing techniques are required for POS tagging that is proficient of drawing interpretations from more delicate dialectal information. From a dialectal outlook, meaning arises from the distinctness between dialectal units, including words, phrases, and so on. These distinctness are of two types: paradigmatic (concerning substitution) and syntagmatic (concerning positioning). To implement Code-Mixed POS tagger all these differences are also needed to be considered.

5. CONCLUSION

The survey shows that in general, various Machine Learning techniques along with POS tagger are used by researchers to implement CM POS taggers for Indian and foreign languages. Much more work is started to perform for code-mixed Indian languages. But an actual tool for code-mixed POS tagging is not yet available on the internet.

REFERENCES

[1] Aravind K. Joshi, "Processing of sentences with intra-sentential code-switching," *Proceedings of the 9th International Conference on Computational Linguistics*, Prague, Czechoslovakia, pp. 145–150, 1982.

[2] John Paolillo, "Language choice on soc. Culture punjab," *Electronic Journal of Communication*, vol. 6(3), 1996.

[3] Thamar Solorio, *et al*., "Overview for the first shared task on language identification in code-switched data," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing 1st Workshop on Computational Approaches to Code Switching*, Doha, Qatar, pp. 62–72, 2014.

[4] Thamar Solorio and Yang Liu, "Part-of-speech tagging for English-Spanish code-switched text," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, pp. 1051–1060, 2008.

[5] Anupam Jamatia, Björn Gambäckand, and Amitava Das, "Part-of-speech tagging code-mixed English-Hindi twitter and facebook chat messages," *Recent Advances in Natural Language Processing (RANLP)*, pp. 239-248, 2015.

[6] Kovida Nelakuditi, Jittadivya Sai, and Radhika Mamidi, "Part-of-Speech Tagging for Code mixed English-Telugu Social media data," *17th International Conference on Intelligent Text Processing and Computational Linguistics Mexico*, 2016.

[7] Kamal Sarkar, "Part-of-Speech Tagging for Code-mixed Indian Social Media Text," *International Conference on Natural Language Processing*, 2015.

[8] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury, "Pos tagging of English-Hindi code-mixed social media content," *In Proceedings of the First Workshop on Codeswitching, EMNLP*, 2014.

[9] Prakash B. Pimpale and Raj Nath Patel, "Experiments with POS Tagging Code-mixed Indian Social Media Text," *NLP Tools Contest on POS Tagging for Code-mixed Indian Social Media Text (POSCMISMT)*, 2015.

[10] R. Sequiera, M. Choudhury, and K. Bali, "POS Tagging of Hindi-English Code Mixed Text from Social Media: Some Machine Learning Experiments," *ICON*, pp. 237-246, Dec 2015.

[11] Kamal Sarkar, "A CRF Based POS Tagger for Code-mixed Indian Social Media Text," *International Conference on Natural Language Processing*, 2016.

[12] Arnav Sharma and Raveesh Motlani, "POS Tagging for Code-Mixed Indian Social Media Text: Systems from IIIT-H for ICON NLP Tools Contest," *International Conference on Natural Language Processing*, Dec 2015.

[13] Ayushman Sisodiya, Donthu Vamsi Krishna, and Sandeep Kumar Begad. "POS Tagging of Code Mixed Text, Project Report," IIT Kharagpur, 2015.

[14] Ajita Singh and Amit Kanskar. "POS Tagging of Hindi-English Code-Mixed Text from Social Media," *International Journal of Science and Research (IJSR)*, vol. 5(10), Oct 2016.

[15] S. Ghosh, S. Ghosh, and D. Das, "Part-of-speech Tagging of Code-Mixed Social Media Text," *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, Austin, TX, pp. 90-97, 2016.

[16] Utsab Barman, Joachim Wagner, and Jennifer Foster, "Part-of-speech Tagging of Code-mixed Social Media Content: Pipeline, Stacking and Joint Modeling," *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, Austin, TX, pp. 30-39, 2016.

[17] Deepak Gupta, Shubham Tripathi, Asif Ekbal, and Pushpak Bhattacharyya, "SMPOST: Parts of Speech Tagger for Code-Mixed Indic SocialMedia Text," *International Conference on Natural Language Processing*, 2016.

[18] Rupal Bhargava, Bapiraju Vamsi Tadikonda, and Yashvardhan Sharma, "BITS_Pilani_Team2@POS Tagging for Code Mixed Indian Social Media," *International Conference on Natural Language Processing*, Dec 2016.

[19] R. Bhargava, R. Bhartia, I. Mishra, and Y. Sharma, "BITS_Pilani_Team1@POS Tagging for Code Mixed Indian Social Media," *International Conference on Natural Language Processing*, Dec 2016.