Proceedings of the WILDRE-6 Workshop @LREC2020, pages 13–17
Marseille, 20 June 2022

Leveraging Sub Label Dependencies in Code Mixed Indian Languages for Part-Of-Speech Tagging using Conditional Random Fields.

Akash Kumar Gautam
Department of Language Science and Technology (LST)
Saarland Informatics Campus, Saarland University, Saarbrücken, Germany
akga00001@stud.uni-saarland.de

Abstract

Code-mixed text sequences often lead to challenges in the task of correct identification of Part-Of-Speech tags. However, lexical dependencies created while alternating between multiple languages can be leveraged to improve the performance of such tasks. Indian languages with rich morphological structure and highly inflected nature provide such an opportunity. In this work, we exploit these sub-label dependencies using conditional random fields (CRFs) by defining feature extraction functions on three distinct language pairs (Hindi-English, Bengali-English, and Telugu-English). Our results demonstrate a significant increase in the tagging performance if the feature extraction functions employ the rich inner structure of such languages.

Keywords: Code-Mixed Text, Indian Languages, Part-Of-Speech Tagging

1. Introduction

In informal settings such as social media, people fluent in multiple languages often converse with each other by changing dialects and languages. This is a highly observable phenomenon among people in India, which is home to several languages. People having text conversations, frequently alternate between a common professional language such as English and other regional languages such as Hindi or Bengali in a single conversation. The primary reason for observing this phenomenon is that in short geo-spatial distances with language diversities, people know neighboring languages as well (Jamatia et al., 2015).

Code-switching has been explored as a research topic in fields such as sociolinguistics, and psycho-linguistics before as well (Joshi, 1982; Paolillo, 1996).

Since code-switching involves alternating between languages below clause level, it leads to creating lexical dependencies which can be leveraged to improve several downstream NLP tasks. In this work, we explore utilizing these sub-label dependencies for improving the part-of-speech (POS) tagging in such a setting.

Current research on POS tagging has concentrated on monolingual text. Hence traditional approaches to this task might not give the best results on specific settings involving code mixed text. To this end, we discuss POS tagging using conditional random fields (CRF) introduced by (Lafferty et al., 2001) in scenarios where there are rich fine-grained sublabels for POS tags.

An example text which demonstrates this scenario is for transliterated Hindi word *achchhaaii*: translation (goodness), which can have multiple levels of tags such as: `ADJ` (adjective) which is the main category followed by subcategories, `QT_QTC`(cardinal quantifier), and `SG` (singular).

In this work, we show that utilizing the labels at multiple levels leads to an improvement in the task of correctly identifying POS tags for the complete text sequence. We achieve this by making use of CRFs, which have the ability to process feature functions given an observation space.

To the best of our knowledge, such an approach of utilizing sub-label dependencies for POS tag identification in code-mixed settings for Indian languages has not been presented before. We present our results on 3 language pairs: Hindi-English, Bengali-English, and Telugu-English. The results of this work indicate that exploiting sub-labels in the text sequences leads to an improvement in the tagging accuracy provided by fine-grained labels.

Contributions: We explain a methodology for defining feature extraction functions leveraging sub-label dependencies based on CRFs along with providing linguistic intuition for using such features in Indian languages (Section 3.). We report the statistical results of our experiments (Section 5.) along with describing various parameter settings (Section 4.) used for the work.

2. Related Work

One of the first approaches for POS tagging of Hindi text was made by (Sangal et al., 1995). Their approach would provide the root form of the word along with a generalized POS category. (Shrivastav et al., 2006) added decision tree-based classification along with this approach to improve the tagging accuracy. (Shrivastava and Bhattacharyya, 2008) made use of a stemmer to create suffixes, which then generated POS tags. Some prior works have also used conditional random fields along with morphological analyzer (Agarwal and Mani, 2006; PVS and Karthik, 2007). Similar attempts were made for Tamil and Bengali (Selvam and Natarajan, 2009; Dhanalakshmi et al., 2008; Ekbal et al., 2007) However, all of these were restricted to monolingual text.

POS tagging for code-mixed text as a research problem is still in its early stage. The earliest attempts made by (Solorio and Liu, 2008a) aimed to make use of machine learning approaches to predict code alternation points for code-mixed English-Spanish data. (Solorio and Liu, 2008b; Bali et al., 2014) used output of language-specific taggers for tagging code-mixed data. (Das and Gambäck, 2015) produced one of the first Indian code-mixed corpora for Hindi-Bengali-English. The traditional approach for automatic identification of such Indian languages utilized n-grams, part-of-speech, lemmas, dictionary-based word classification (Barman et al., 2014a; Barman et al., 2014b; Bali et al., 2014)

3. Methods

Given a sequence of tokens in a sentence consisting of x = (x_1, ... $x_{|x|}$) and the relevant POS tags as, y = (y_1, ... $y_{|x|}$), the CRF model (Lafferty et al., 2001) is considered as:

$$p(y \mid x; w) \propto \prod_{i=n}^{|x|} exp(w.\phi(y_{i-n}, ...y_i, x, i)) \quad (1)$$

Here, n defines the model order, w is the model parameter, and ϕ is the feature extraction function. Each $y_i \in Y$ for $i \in 1$ ... $|x|$, denotes the tag set. In the next sections, we describe the feature functions which can model sequence-based dependencies for code-mixed text. The baseline features define a naive set of functions that associate the relationship between the POS tag label and the token. Expanded features utilize the sub-label dependencies by exploiting the inner structure of fine-grained labels.

3.1. Baseline Feature Set

Based on work of (Ratnaparkhi and others, 1996; Silfverberg et al., 2014), the baseline features associates a set of functions for a word form x_i with y_i (label), where i is it's position in the sequence. These functions are:

- Bias, true irrespective of the input word-form.
- Word forms x_{i-2}, .., x_{i+2} for given x_i, including the length.
- Language of the current word form x_i.
- Prefix and suffix of the current word form of various lengths upto δ = 4.
- Presence of url, user-mentions, hashtags in x_i, assigned by a boolean value.
- Boolean function indicating, if the word form x_i is an upper capital string or is a number.

These serve a practical purpose in Indian languages where case (nominative, accusative, genitive), number (singular, plural), and gender (masculine, feminine) are inflected through suffix and prefix in word-forms (Schmid and Laws, 2008). Most Indian languages follow case-based dependent marking. For the current task, these are defined for a word-form only if its length is more than 4. Capitalization of a word-form helps in identification whenever a token is used as a proper noun (Silfverberg et al., 2014). Hence, we can say that the mentioned feature functions are representative of the data highly prevalent on social media platforms and have the ability to capture sequence-based dependencies in code-mixed settings.

3.2. Expanded Feature Set

In this section, we describe the expanded feature set which has the ability to model the sub-label dependencies in a given sequence. Fine-grained labels include multiple levels of labeling for POS tags (sub-labels) which are used to indicate the main category of the token, followed by its sub-category. Such labels are known as compound labels.

Instead of associating feature functions for a word-form x_i with just label y_i, we partition any compound label into its sub-components (s). As an example, consider the Hindi word-form *tha*: translation (was), consisting of the compound label, { `V + VAUX` }, hence listing that this word-form has the main category as a verb, and within the given utterance, it occurs as an auxiliary verb.

Let S be the set of all sub-label components for a compound label. Then, we individually associate feature functions with all the sub-labels such that $s \in S$. We describe the process of partitioning a compound label in detail in section 4.2.. This approach aims to utilize the morphological rich structure of highly inflected Indian languages for improving the tagging accuracy.

3.3. Linguistic Motivation For Expanded Feature Set

This section aims to provide linguistic intuition behind selecting the mentioned expanded features and why leveraging sub-label dependencies for a token provides a better representation of a sequence for Indian languages.

Consider a noun based transliterated word-form in Hindi: *nadiya* (`NOUN`): translation (river – plural). For such a word-form, the baseline feature set would just associate 2-suffix *-ya* to the compound label { `NOUN + PLURAL` }. In Hindi, morpheme -`ya` is used as a suffix based marker for plural.

The expanded feature set on the other hand would associate the 2-suffix *-ya* to both the main label `NOUN`, and the sub-label `PLURAL` individually. Such an approach of distribution of labels would be useful for correct identification of a different verb-based word-form in Hindi, *shaktiya* (`VERB`): translation (power – plural) which is also formed by inflecting the 2-suffix morpheme *-ya* to the root word, *shakti*.

4. Experiments

In this section we describe constituents for the experiments, including data, tag-set and partitioning of labels for the expanded feature set.

4.1. Data

We use the dataset provided by (Jamatia et al., 2015) for the mentioned research problem. It contains text conversations recorded from social media platforms such as Twitter, WhatsApp, and Facebook, code-mixed in these language pairs: Hindi-English, Bengali-English, and Telugu-English. The mentioned conversations are labeled into appropriate fine-grained POS tags along with the language of each token in the utterance. Please refer to table 2 for an overview of the number of utterances for each language pair in the dataset.

[user]	why	not	hike	the	petrol	price	to	120	rs/Ltr	...	,	baar	baar	shock
@	QT_QTC	RP_NEG	V_VM	DT	JJ	N_NN	RP_RPD	$	N_NN	RD_PUNC		N_NNP	RP_RPD	V_VM
dene	se	accha	hai	ki	ek	baar	mein	hi	de	diya	jaye	!		
N_NNV	PSP	RB_AMN	V_AUX	PSP	QT_QTO	RP_RPD	PSP	RP_RPD	V_VM	V_VM	V_VAUX	RD_PUNC		

Table 1: Sample sentence from the dataset (Jamatia et al., 2015) with code-mixed Hindi-English text and fine-grained POS tag labels. Original utterances in the dataset includes Hindi words as transliterated text.

Language Pairs	Hindi-English	Bengali-English	Telugu-English
#Utterances	2630	624	1279

Table 2: Total number of text utterances for each language pair in the dataset (Jamatia et al., 2015).

POS-tags were assigned to each token by manual annotation with substantial agreement over the labels after deciding the utterance boundary. Labels over text conversations use tagset introduced by (Gimpel et al., 2010) for Twitter-specific data and a set of POS tags for Indian languages (Jha et al., 2009) for a fine-grained annotation scheme. Each instance of the datapoint includes the token, identified language for a token, and labeled POS tag. There are dedicated tags for identifying universal acronyms or punctuations as tokens in the dataset. Table 1 shows a sample sentence from the dataset with code mixed Hindi-English text and POS tag labels. Personally identifiable information for a social media user has been removed from the example presented.
The authors of the dataset mention that even though corpus is bi-lingual, there might be occasional instances of tri-quad-lingual mix in a single utterance as well. For each language pair, the total number of utterances were split into the ratio of 80:10:10 as train, test, and validation splits respectively.

4.2. Partitioning of Labels

In this section, we describe the process of splitting the compound labels mentioned in section 3.2. for an expanded feature set. A fine-grained annotation scheme for POS tags mentioned in (Jamatia et al., 2015) focuses on identifying the main category of the token, followed by a descriptive sub-category. Our distribution process aims to leverage that.
For example, given a compound label (V_VM) for a word-form, it is split in a way such that it identifies the main category of the token as (verb) and the sub-category of the token as a (main verb), hence such a label would be distributed into the set {V, VM}. Compound labels in the dataset for a token are identified by the presence of underscore (_) within a POS tag for a token. Not every label for a token in the dataset is a compound label. The described splitting scheme was followed before performing the experiments, hence they were not optimized through the development set.

4.3. Model Specifications

For the code-mixed settings for Indian languages, we explore the baseline feature set and the expanded feature set for first-order ($n = 1$) and second-order ($n = 2$) CRF models. The CRF model parameters in all the cases were estimated using Averaged Perceptron algorithm (Collins, 2002). We use sklearn crf-suite [1] open-source implementation for this work. The maximum number of iterations for the training algorithm was set to 100. The parameters were evaluated on the validation set, with the best-performing ones finally applied to the test set. Instances of the test set were decoded using the Viterbi algorithm.

5. Results and Discussions

Sub-Label Dependencies: Table 3 summarizes the weighted F1 scores of the baseline feature set and expanded feature set for the first-order and second-order CRF models for the 3 language pairs in the dataset. Compared to the standard baseline features, the expanded features show an improvement for all the language pairs, for both first and second-order models. These results are in line with the linguistic intuition for using sub-label dependencies for Indian languages.

Model Order: Another interesting observation is the increased weighted F1-score for first-order models with the expanded feature set, compared to baseline features for the second-order models. However, within the experiments performed, this is observed only for 2 language pair: Hindi-English, and Bengali-English. This suggests that as opposed to increasing the model order, utilizing sub-label dependencies for Indian languages might lead to a better improvement of results. The best set of results for all the language pairs was obtained using an expanded feature set within the second-order model. Tagging accuracy percentage of the models follows the same trend, as described previously for weighted F1 scores, however, for brevity, these have been omitted.

Feature Ablation: In table 4 we report the effects of individual features on the second-order CRF model for Hindi-English language pair on the expanded feature set. The performance is reported in terms of tagging accuracy percentage. From the table, it can be concluded that adding prefixes and suffixes of varying lengths (δ) to the feature extraction function leads to a considerable increment in the performance of the model. Finally identifying URLs, mentions and capitalization help improve the performance most for the text data, as these are efficiently able to capture sequence-based dependencies for social media text.

6. Conclusion

In this work, we evaluate the ability to utilize sub-label dependencies in Indian languages for improving the tagging accuracy of the code-mixed text. We analyze results

[1] https://sklearn-crfsuite.readthedocs.io/en/latest/api.html

Language-Pairs	**First Order** ($n = 1$) Baseline Features	Expanded Features	**Second Order** ($n = 2$) Baseline Features	Expanded Features
Hindi - English	0.70	0.77	0.71	**0.81**
Bengali - English	0.65	0.73	0.69	**0.83**
Telugu - English	0.70	0.71	0.71	**0.72**

Table 3: Table showing comparison between baseline feature set and expanded feature set for first order and second order CRF models explored for all language pairs through **weighted F1-score**. Best results are highlighted in **bold**.

Features	Accuracy %
Current word-form (x_i)	74.16
+ Language and Length	75.10
+ Prefix-Suffix (δ = 1)	76.81
+ Prefix-Suffix (δ = 2)	77.91
+ Prefix-Suffix (δ = 4)	78.87
+ Urls, mentions, capitalization	80.09

Table 4: Feature ablation for second order model (expanded feature set) on Hindi-English language pair. Performance measured through **tagging accuracy percentage.**

in three different language pairs: Hindi-English, Bengali-English, and Telugu-English over first and second-order CRF models. Preliminary conclusions from the results show a step in the right direction. We observe that expanded feature set making use of sub-label dependencies shows a vast improvement against the baseline.
In the future, we aim to utilize neural network architectures like LSTM's having the ability to process lexical sequences over feature functions defined by sub-label dependencies. Another direction to take this research could be to evaluate the performance by having a different splitting criterion for a compound label as opposed to the one described in this paper.

7. Acknowledgments

We would like to thank Dr. Alexander Koller for his initial suggestions regarding research in this paper. We also extend our gratitude to anonymous reviewers for their insightful comments on this work.

8. Bibliographical References

Agarwal, H. and Mani, A. (2006). Part of speech tagging and chunking with conditional random fields. In *the Proceedings of NWAI workshop*.

Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (2014). "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126.

Barman, U., Das, A., Wagner, J., and Foster, J. (2014a). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Barman, U., Wagner, J., Chrupała, G., and Foster, J. (2014b). Dcu-uvt: Word-level language classification with code-mixed data. In *Proceedings of the first workshop on computational approaches to code switching*, pages 127–132.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002)*, pages 1–8.

Das, A. and Gambäck, B. (2015). Code-mixing in social media text: the last language identification frontier?

Dhanalakshmi, V., Anandkumar, M., Vijaya, M., Loganathan, R., Soman, K., and Rajendran, S. (2008). Tamil part-of-speech tagger based on svmtool. In *Proceedings of the COLIPS International Conference on Asian Language Processing*, pages 59–64.

Ekbal, A., Mondal, S., and Bandyopadhyay, S. (2007). Pos tagging using hmm and rule-based chunking. *The Proceedings of SPSAL*, 8(1):25–28.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2010). Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Jamatia, A., Gambäck, B., and Das, A. (2015). Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. Association for Computational Linguistics.

Jha, G. N., Gopal, M., and Mishra, D. (2009). Annotating sanskrit corpus: adapting il-posts. In *Language and Technology Conference*, pages 371–379. Springer.

Joshi, A. (1982). Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Paolillo, J. C. (1996). Language choice on soc. culture. punjab. *Electronic Journal of Communication/La revue électronique de communication*, 6(3).

PVS, A. and Karthik, G. (2007). Part-of-speech tagging and chunking using conditional random fields and transformation based learning. *Shallow Parsing for South Asian Languages*, 21:21–24.

Ratnaparkhi, A. et al. (1996). A maximum entropy model for part-of-speech tagging. In *EMNLP*, volume 1, pages 133–142. Citeseer.

Sangal, R., Chaitanya, V., and Bharati, A. (1995). *Natu-*

ral language processing: a Paninian perspective. PHI Learning Pvt. Ltd.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784.

Selvam, M. and Natarajan, A. (2009). Improvement of rule based morphological analysis and pos tagging in tamil language via projection and induction techniques. *International journal of computers*, 3(4):357–367.

Shrivastav, M., Melz, R., Singh, S., Gupta, K., and Bhattacharyya, P. (2006). Conditional random field based pos tagger for hindi. *Proceedings of the MSPIL*, pages 63–68.

Shrivastava, M. and Bhattacharyya, P. (2008). Hindi pos tagger using naive stemming: harnessing morphological information without extensive linguistic knowledge. In *International Conference on NLP (ICON08), Pune, India*.

Silfverberg, M., Ruokolainen, T., Lindén, K., and Kurimo, M. (2014). Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–264.

Solorio, T. and Liu, Y. (2008a). Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981.

Solorio, T. and Liu, Y. (2008b). Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060.