Creation of Corpus and Analysis in Code-Mixed Kannada-English Social Media Data for POS Tagging

Appidi Abhinav Reddy, Vamshi Krishna Srirangam,
Suhas Darsi and Manish Shrivastava
Language Technologies Research Centre (LTRC)
Kohli Centre on Intelligent Systems(KCIS)
International Institute of Information Technology, Hyderabad, India.
(abhinav.appidi, v.srirangam, darsi.suhas)@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Part-of-Speech (POS) is one of the essential tasks for many Natural Language Processing (NLP) applications. There has been a significant amount of work done in POS tagging for resource-rich languages. POS tagging is an essential phase of text analysis in understanding the semantics and context of language. These tags are useful for higher-level tasks such as building parse trees, which can be used for Named Entity Recognition, Coreference resolution, Sentiment Analysis, and Question Answering. There has been work done on code-mixed social media corpus but not on POS tagging of Kannada-English code-mixed data. Here, we present Kannada-English code-mixed social media corpus annotated with corresponding POS tags. We also experimented with machine learning classification models CRF, Bi-LSTM, and Bi-LSTM-CRF models on our corpus.

1 Introduction

The advent of social media like Twitter, Facebook, and Reddit has accelerated the communication between people of all colors, nations, and languages. Though the platform exists, the barriers to communication still exist due to languages. Many researchers are trying to solve this through various methods.

India is a land of multiple languages and majority of people are multilingual and tend to mix words from different languages in written text and also in speech. This interchanging language method involves complex grammar and is commonly addressed by terms 'Code-mixing' and 'Code-switching' as described by Lipski (1978). Code-switching refers to the use of words or phrases from different languages within the same speech context, whereas Code-mixing refers to the use of words or phrases from different languages in the same sentence. We can understand the difference between code-mixing and code-switching from the positions of altered elements. Code-mixing refers to the intra-sentential modification of codes, whereas code-switching refers to the inter-sentential modification of codes.

1.1 Characteristic of Code-Mixed Kannada-English Data

As explained above mixing happens at phrase, word, syntactic and morphological level too. Following are few more examples :

1. **Morphological level:** The word 'cinemagalu' in Kannada, the root word 'cinema' is borrowed from English and 'galu' is a Kannada morphene that marks plurality.
2. **Phrase level:** This is a completely code-mixed sentence. For example, 'Kelsa bittu pitch reporter aagu olle future ide!' which means 'Leave your work and become pitch reporter, you have great future in that!'. Here the statement follows the structure of Kannada with English words embedded in it.
3. **Word level:** This is language mixing occuring at word level. A complete word from English language is taken into Kannada language. An example: 'Ee thara branch ideya' which means 'Is there a branch like this?'.
4. **Syntactic level:** There are occurrences in Kannada-English CM data where inter-sentential mixing takes place. For example, 'Born and brought up in bengaluru, Yaako nange mysoor thumba ista, mysoor alli kelsa sikdre ready to shift.'

While there are robust solutions currently to handle non-code-mixed data, the same is not true for code-mixed data. One of the keys to solving any

Proceedings of the 17th International Conference on Natural Language Processing, pages 101–107
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

higher-level NLP tasks is to do POS tagging. While POS tagging on English is very mature at this point, POS tagging for code-mixed in low-resource languages is relatively uncommon. In this paper, we have tried to address this problem. Here, we present Kannada-English code-mixed social media corpus annotated with corresponding POS tags.

Due to unstructured, informal, and incomplete information available in the data, it complicates the task of Code-mixed Kannada-English. Following are the challenges associated with the corpus.

- **Ambiguous words**: A word in one language can have a different meaning in other languages. For example, the word 'Bali' in English, which is a place in Indonesia, also used in Kannada which means 'Near'.

- **Word-level Code-mixing**: In the word 'Kanglish', its a fusion of two words Kannada and English at word level. This is similar to language mixing at word-level.

- **Word Orders**: English and Indian languages follow different word orders. Indian languages follow Subect-Object-Verb format, whereas English language follows Subject-Verb-Object format.

- **Reduplication**: People tend to use a second word with first word, which does not have a meaning on its own. The second word when addressed together with the first word it becomes a multi word expression. For example 'postu geestu', 'desha gesha', 'man ban'.

- **Variable Lexical Representations**: Users on social media have preference for their own way of native words like for example 'hogilla' is a Kannada word and it can be written as 'hogila', 'hgilla' etc.

Here are an instance depicting Kannada-English code-mixed nature and its translation.

T1 : *"@Suharsh2512 oho, idyaavdo brilliant facility. Nanna phone alli sound barutte.. ondond sala baralla. Hyaage nodu..."*
Translation: *"@Suharsh212 Oho...this is some brilliant facility...in my phone there is sound..once there is no sound...see how it is "*

2 Background and Related Work

POS tagging is a crucial stage in the NLP pipeline (Cutting et al., 1992) and has been explored extensively by Toutanova et al. (2003). Gimpel et al. (2010) and Owoputi et al. (2013) worked on the POS tagging of social media data. POS tagging for English using Dynamic Feature Induction with an accuracy of 97.64% was done on the WallStreet journal data set by Choi (2016).

POS tagging work has been done on Indian monolingual languages. Earlier work in POS tagging for Indian languages was mainly based on rule-based approaches (Antony and Soman, 2011). Some works in POS tagger system in Hindi done by Singh et al. (2006) and in the Bengali language was done by Ekbal et al. (2009) and in Telugu by RamaSree and Kusuma Kumari (2007).

Not many works were done on the POS tagger on Code-mixed data. POS taggers have been trained on Hindi-English code-mixed posts generated on Facebook (Vyas et al., 2014; Sharma et al., 2016). Only one public dataset of English-Hindi code-mixed Twitter posts annotated for POS tags exists (Jamatia and Das, 2016). Some of the recent works in code-mixed includes POS on code-mixed Telugu-English by Nelakuditi et al. (2016) and in NER in Telugu-English code-mixed social media data by Srirangam et al. (2019).

There are not many works done on Kannada because of the scarcity of quality annotated data. Recent works in POS tagging on Kannada were experimented only with traditional ML techniques like HMM, CRF, or SVM (BR and Kumar, 2012; Antony and Soman, 2010). Todi et al. (2018) built a Kannada POS tagger using machine learning and neural network models.

There have been very few works done on Kannada-English code-mixed data. Lakshmi and Shambhavi (2017) presented an automatic language identification system for code-mixed Kannada-English Social media text. Shalini et al. (2018) worked on sentiment analysis for Code-Mixed Kannada-English Social Media Text.

To the best of our knowledge, the corpus created for this paper is the first ever Kannada-English code-mixed social media corpus with POS tags.

3 Corpus Creation and Annotation

This corpus consists of Kannada-English code mixed tweets scraped from Twitter for the past six years based on topics such as sports, trending hashtags, politics, movies, events, and others not limited to a particular domain. The tweets were collected using twintproject[1]-an opensource twitter intelligence tool. We retrieved over 318,000 tweets using the mentioned tool. After extensive cleaning and pre-processing of tweets, we were left with 6468 code-mixed Kannada-English tweets. We have done extensive pre-processing of tweets and retrieved them in JSON format. This JSON formatted data includes metadata like URLs, usernames, retweets, tweet IDs, likes, full names, and others.

The following steps were followed during pre-processing :

- Removing useless, noisy tweets, i.e., tweets containing only hashtags and URLS.
- Tweets that were written in only English or only Kannada were removed too.
- Tweets that having a minimum of ten words and contain linguistic units from both English and Kannada are only considered.
- Tweet Tokenizer is used to do Tokenisation of tweets.

The corpus will be made available for public use as soon as possible. The following explains the mapping of the tokens with their respective tags.

3.1 Annotation: Parts of Speech

Since the paper focuses on two different languages Kannada and English, we follow the Universal POS proposed by Petrov et al. (2011), which covers POS tags across all languages. There are 17 tags in the Universal POS[2], which we are following such as adjectives(ADJ), adposition(ADP), adverb(ADV), auxiliary(AUX), coordinating conjunction(CCONJ), determiner(DET), interjection(INTJ), noun(NOUN), numeral(NUM), particle(PART), pronoun(PRON), proper noun(PROPN), punctuation(PUNCT), subordinating conjunction(SCONJ), symbol(SYM), verb(VERB), and other(X). These tags are used in the annotation of our corpus. 'X' tag in the Universal POS is used to denote typos, foreign words, unknown abbreviations, and others. We included punctuation symbols under the category 'PUNC'. Following is an example of an annotated tweet and its translation.

Tag	Cohen Kappa	Tokens
ADJ	0.84	6209
ADP	0.85	7000
ADV	0.85	11765
AUX	0.92	2098
CCONJ	0.83	2252
DET	0.88	3334
INTJ	0.87	943
NOUN	0.91	44533
NUM	0.92	1220
PART	0.89	569
PRON	0.89	17549
PROPN	0.91	7411
PUNCT	0.90	15602
SCONJ	0.82	1713
SYM	0.83	617
VERB	0.81	32545
X	0.85	1381

Table 1: Inter Annotator Agreement.

T2 : *"Haha/INTJ ashtu/ADV idea/NOUN illade/ADV gowdru/NOUN bengaluru/NOUN north/NOUN bittu/VERB tumukur/NOUN hogilla/ADV"*
Translation: "Haha without having much idea gowda left bengaluru north and went to tumukur."

3.2 Inter-annotator Agreement

Two people who are with linguistic backgrounds, both proficient in Kannada and English, manually did the annotations of the POS tags. Inter Annotator Agreement (IAA) is used to validate the quality of the annotation between two annotation sets of 6468 tweets and 156761 tokens using Cohen's Kappa coefficient (Hallgren, 2012) (refer Table 1 for Score). The agreement is significantly high.

4 Corpus Statistics

We have collected more than 318,000 of tweets from Twitter using TwintProject. After extensive cleaning, we were left with 6468 code-mixed Kannada-English tweets, as part of annotation using sixteen POS tags along with 'X' tag for foreign words, we tagged 156761 tokens (refer Table 1). We made sure that all the words in the corpus are in Roman script. We used hashtags related to sports, trending hashtags, politics, movies, events, and others in collecting the corpus.

[1] https://github.com/twintproject/twint

[2] https://universaldependencies.org/u/pos/

5 Experiments

We present the experiments using a combination of features and systems. To understand the effect of different parameters and features of the model, we performed several experiments. With some set of features at once and all at a time simultaneously, we performed experiments while changing the parameters of the model, like regularization parameters and algorithms of optimization like 'L2 regularization', 'Average Perceptron'and 'Passive Aggressive' for CRF, optimization algorithms and loss functions in LSTM. We used three-fold cross-validation for CRF. We used 'scikit-learn,' 'Tensorflow,' and 'Keras' libraries to implement the above algorithms.

5.1 Conditional Random Field (CRF)

CRFs are type of discriminative undirected probabilistic graphical model. In natural language processing, linear chain CRFs are popular, which implement sequential dependencies in predictions.[3] It is a supervised learning method and most often used for structured prediction tasks. In CRF, a set of feature functions are defined to extract features for each word in a sentence. It has applications in NER, POS tagging, among others. When it comes to POS tagging, it has been proven to be better than the tree-based models.

5.2 LSTM

Long Short Term Memory (LSTM) is a special kind of RNN architecture that is well suited for classification and making predictions based on time series data. LSTMs are capable of capturing only past information. In order to overcome this limitation Bidirectional LSTMs are proposed where two LSTM networks run in forward and backward directions capturing the context in either directions.

5.3 LSTM-CRF

The Bi-LSTM-CRF is a combination of bidirectional LSTM and CRF (Huang et al., 2015; Lample et al., 2016). The Bi-LSTM model can be combined with CRF to enhance recognition accuracy. This combined model of Bi-LSTM-CRF inherits the ability to learn past and future context features from the Bi-LSTM model and use sentence-level tags to predict possible tags using the CRF layer. Bi-LSTM-CRF has been proved to be a powerful model for sequence labeling tasks like POS tagging, shallow parsing, and NER.

5.4 Features

The features to our machine learning models consist of lexical, word-level and character features such as char N-Grams of size 2 and 3 in order to capture the information from emojis, mentions, suffixes in social media like '#,' '@,' numbers in the string, numbers, punctuation. Features from adjacent tokens are used as contextual features.

1. **Character N-Grams:** Character N-Grams are proven to be efficient in the task of classification of text and are language-independent (Majumder et al., 2002). They are helpful when there are misspellings in the text (Cavnar et al., 1994; Huffman, 1995; Lodhi et al., 2002). Group of chars can help in capturing the semantic information. Character N-Grams are especially helpful in cases like code mixed language where there is free use of words, which vary significantly from the standard Kannada-English words.

2. **Word N-Grams:** Bag of words has been a staple for languages other than English (Jahangir et al., 2012) in tasks like NER and POS. Thus, we use adjacent words as a feature vector to train our model as our word N-Grams. These are also called contextual features. We used Word N-Grams of size 3 in the paper.

3. **Common Symbols:** It is observed that currency symbols, brackets like '(,' '[,' etc. And other symbols are followed by numeric or some mention, are present in the corpus which direct to symbol tag under Universal POS. Hence, the presence of these symbols is a good indicator of the words before or after them for being a 'SYM' tag in POS tagging.

4. **Numbers in String:** In social media, we see people using alphanumeric characters, generally to save the typing effort, to showcase their style or shorten the message length. When observed in our corpus, words containing alphanumeric are generally tagged under 'NUM' tag.

5. **Mentions and Hashtags:** People use '@' mentions to refer to persons or organizations, they use '#' hashtags in order to make something notable or to make a topic trending.

[3] https://en.wikipedia.org/wiki/Conditional$_{r}andom_{f}ield$

Thus the presence of these two gives a reasonable probability for the word being a named entity which counts under proper nouns.

6. **Capitalization:** In social media, people tend to use capital letters to refer to the names of persons, organizations and persons; at times, they write the entire name in capitals (Von Däniken and Cieliebak, 2017) to give particular importance or to denote aggression. This gives rise to a couple of binary features. One feature is to indicate if the beginning letter of a word is capitalized, and the other is to indicate if the entire word is capitalized.

6 Results and Discussion

Table 2 shows CRF results with 'l2-sgd' (Stochastic Gradient Descent with L2 regularization) algorithm for 200 iterations. The c2 value in the CRF model refers to the 'L2 regression'. Experiments using the algorithms 'pa' (Passive-Aggressive) and 'ap' (Averaged Perceptron) resulted in similar F1-scores of 0.79. The table 3 shows results after removing each particular feature. Example prediction of our CRF model is shown under appendix section.

In both the experiments Bi-LSTM and Bi-LSTM-CRF, we experimented with the optimizer, activation functions, and the number of epochs. After several experiments, the best result we came through was using 'softmax' as activation function, 'rmsprop' as an optimizer and 'categorical cross-entropy' as our loss function. Table2 shows the results of BiLSTM on our corpus using thirty epochs, and also shows the results of Bi-LSTM-CRF on our corpus using twenty epochs, both with random initialization of embedding vectors. The training, validation, and testing for both experiments are 60%, 10%, and 30% of the total data, respectively. Bi-LSTM resulted in best F1-score of 0.80 and Bi-LSTM-CRF with best F1-score of 0.81.

Tag	CRF	Bi-LSTM	BiL-CRF
ADJ	0.58	0.52	0.58
ADP	0.75	0.73	0.78
ADV	0.75	0.79	0.72
AUX	0.99	1.00	0.99
CCONJ	0.99	0.31	0.99
DET	0.85	0.74	0.88
INTJ	0.97	0.93	0.87
NOUN	0.83	0.84	0.84
NUM	0.69	0.76	0.74
PART	1.00	0.99	1.00
PRON	0.67	0.60	0.63
PROPN	0.86	0.77	0.77
PUNCT	1.00	1.00	1.00
SCONJ	0.77	1.00	0.75
SYM	0.80	0.74	0.78
VERB	0.70	0.70	0.69
X	0.79	0.80	0.81
weighted avg	0.80	0.79	0.79

Table 2: Table shows F1-scores for CRF, Bi-LSTM and Bi-LSTM-CRF respectively.

Feature removed	Precision	Recall	F1
Char N-Grams	0.66	0.50	0.45
Word N-Grams	0.62	0.53	0.50
Common Symbols	0.66	0.55	0.52
Numbers in String	0.62	0.56	0.55
Mentions, Hashtags	0.60	0.56	0.54
Capitalization	0.59	0.55	0.53

Table 3: Feature(removed) Specific Results for CRF.

7 Conclusion and Future Work

Our Contributions are as follows:

1. Presented an annotated Kannada-English code-mixed corpus for POS, which is, to the best of our knowledge is the first ever corpus. The corpus will be made available online.

2. We have experimented with the machine learning models CRF, Bi-LSTM, and Bi-LSTM-CRF on our data, the F1-score for which is 0.79, 0.80, and 0.81 respectively, which looks good considering the amount of research done in this new area.

3. We are introducing and addressing Part-of-Speech of code-mixed Kannada-English data as a research problem.

For future work, the corpus can also be enriched by giving the NER tags for each token. The size of the corpus can be increased with more data. The problem can be adapted for POS tagging in multilingual code-mixed data.

References

PJ Antony and KP Soman. 2010. Kernel based part of speech tagger for kannada. In *2010 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2139–2144. IEEE.

PJ Antony and KP Soman. 2011. Parts of speech tagging for indian languages: a literature survey. *International Journal of Computer Applications*, 34(8):0975–8887.

Shambhavi BR and Ramakanth Kumar. 2012. Kannada part-of-speech tagging with probabilistic classifiers. *international journal of computer applications*, 48(17):26–30.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer.

Jinho D Choi. 2016. Dynamic feature induction: The last gist to the state-of-the-art. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 271–281.

Douglass Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing*, pages 133–140.

Asif Ekbal, Md Hasanuzzaman, and Sivaji Bandyopadhyay. 2009. Voted approach for part of speech tagging in bengali. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 120–129.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2010. Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Stephen Huffman. 1995. Acquaintance: Language-independent document categorization by N-grams. Technical report, DEPARTMENT OF DEFENSE FORT GEORGE G MEADE MD.

Faryal Jahangir, Waqas Anwar, Usama Ijaz Bajwa, and Xuan Wang. 2012. N-gram and gazetteer list based named entity recognition for Urdu: A scarce resourced language. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 95–104.

Anupam Jamatia and Amitava Das. 2016. Task report: Tool contest on pos tagging for code-mixed indian social media (facebook, twitter, and whatsapp) text@ icon 2016.". *Proceedings of ICON*.

BS Sowmya Lakshmi and BR Shambhavi. 2017. An automatic language identification system for code-mixed english-kannada social media text. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pages 1–5. IEEE.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

John Lipski. 1978. Code-switching and the problem of bilingual competence. *Aspects of bilingualism*, 250:264.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444.

P Majumder, M Mitra, and BB Chaudhuri. 2002. N-gram: a language independent approach to IR and NLP. In *International conference on universal knowledge and language*.

Kovida Nelakuditi, Divya Sai Jitta, and Radhika Mamidi. 2016. Part-of-speech tagging for code mixed english-telugu social media data. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 332–342. Springer.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

RJ RamaSree and P Kusuma Kumari. 2007. Combining pos taggers for improved accuracy to create telugu annotated texts for information retrieval. *Dept. of Telugu Studies, Tirupathi, India*.

K Shalini, HB Barathi Ganesh, M Anand Kumar, and KP Soman. 2018. Sentiment analysis for code-mixed indian social media text with distributed representation. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1126–1131. IEEE.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Srivastava, Radhika Mamidi, and Dipti M Sharma. 2016. Shallow parsing pipeline for hindi-english code-mixed social media text. *arXiv preprint arXiv:1604.03136*.

Smriti Singh, Kuhoo Gupta, Manish Shrivastava, and Pushpak Bhattacharyya. 2006. Morphological richness offsets resource demand–experiences in constructing a pos tagger for hindi. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 779–786.

Vamshi Krishna Srirangam, Appidi Abhinav Reddy, Vinay Singh, and Manish Shrivastava. 2019. Corpus creation and analysis for named entity recognition in telugu-english code-mixed social media data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 183–189.

Ketan Kumar Todi, Pruthwik Mishra, and Dipti Misra Sharma. 2018. Building a kannada pos tagger using machine learning and neural network models. *arXiv preprint arXiv:1808.03175*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for Computational Linguistics.

Pius Von Däniken and Mark Cieliebak. 2017. Transfer learning and sentence level features for named entity recognition on tweets. In *3rd Workshop on Noisy User-generated Text (W-NUT), Copenhagen, 7 September 2017*, volume 3, pages 166–171. ACL.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979.

A Appendices

A.1 Example Prediction of CRF

Word	Truth	Predicted
Haha	INTJ	INTJ
ashtu	ADV	VERB
idea	NOUN	NOUN
illade	ADV	VERB
gowdru	NOUN	NOUN
bengaluru	NOUN	NOUN
north	NOUN	NOUN
bittu	VERB	NOUN
tumukur	NOUN	NOUN
hogilla	ADV	ADV