# Part-of-Speech Tagger for Konkani-English Code-Mixed Social Media Text

Akshata Phadte(✉) and Radhiya Arsekar(✉)

Department of Computer Science and Technology, Goa University,
Taleigao Plateau, Goa, India
akshataph07@gmail.com, radhiya.arsekar@gmail.com

**Abstract.** In this paper, we propose efficient and less resource-intensive strategies for Konkani-English code-mixed social media text. which witnesses several challenges as compared to tagging general normal text. Part-of-Speech Tagging is a primary and an important step for many Natural Language Processing Applications. This paper reports work on annotating code-mixed Konkani-English data collected from social media site Facebook, which consists of more than four thousands posts from Facebook and developed automatic Part-of-Speech Taggers for this corpus. Part-of-Speech tagging is considered as a classification problem and we use different classifiers such as CRFs, SVM with different combinations of features.

**Keywords:** Code-mixing · Social media text · Part-of-Speech tagging

## 1 Introduction

India is a land of many languages. People on social media often use more than one language to express themselves. But the problem starts with Multilingual speakers tend to exhibit code-mixing and code-switching in their use of language on social media platforms. Code-mixing refers to the mixing of two or more languages or language varieties in speech. Code-mixing occurs due to various reasons. According to a work by [1], the major reasons for Code-Mixing:- 45% Real lexical needs, 40% Talking about a particular topic and 5% For content clarification. [3] noted that the complexity in analyzing code-mixed social media text (CMST) stems from non adherence to a formal grammar, spelling variations, lack of annotated data, inherent conversational nature of the text and of-course, code-mixing. Therefore, there is a need to create datasets and Natural Language Processing tools for code-mixed social media text as traditional tools are ill-equipped for it. Taking a step in this direction, we present our work on building a POS tagger for Konkani-English code-mixed data collected from social media site Facebook.

M. Silberztein et al. (Eds.): NLDB 2018, LNCS 10859, pp. 303–307, 2018.
https://doi.org/10.1007/978-3-319-91947-8_31

## 2 Related Work

Code-mixing being a relatively newer phenomena has gained attention of researchers only in the past two decades. POS taggers on monolingual data give an accuracy of about 97.3% for English text [6]. They are often seen as sequence labeling problems and have used the context based information in the form of lexical and sub-lexical characteristics of neighboring words. But in code-mixed setting, the context information can be in a different language which makes the understanding difficult. [3] reported challenges in processing Hindi-English CMST and performed initial experiments on POS tagging. Their POS tagger accuracy fell by 14% to 65% without using gold language labels and normalization. Thus, language identification and normalization are critical for POS tagging [3]. [7] also built a POS tagger for Hindi-English CMST using Random Forests on 2,583 utterances with gold language labels and achieved an accuracy of 79.8%.

[8] further improved this POS tagger, increasing the accuracy to 93%. [11] worked on a complete pipeline for shallow parsing and performed tokenisation, language identification, normalisation, POS tagging and finally, shallow parsing and achieved accuracy of 83.4% for code-mixed Hindi-English social media text.

## 3 Data Preparation

Significant studies and dataset of the code-mixing phenomenon can be found in [2]. These works discuss the dataset preparation and dataset statistics of code-mixing of Konkani-English as well as its linguistic nature. For the POS tagging of Konkani-English language we extracted the code-mixed corpus which was discussed in [2]. We then manually tagged them by their language, normalisation form and by their POS tags.

### 3.1 Dataset Annotation Guidelines

The creation of this linguistic resource involved Language identification, Normalisation and POS tagger layer. The following paragraphs describe the annotation guidelines for these tasks in detail.

1. **Language Identification:** Every word was given a tag out of three - en, kn and rest to mark its language. Words that a bilingual speaker could identify as belonging to either Konkani or English were marked as ‘kn’ or ‘en’, respectively. The label ‘rest’ was given to symbols, emoticons, punctuation, named entities, acronyms and foreign words.
2. **Normalisation:** Words with language tag ‘kn’ in Roman script were labeled with their standard form in the native script of Konkani Devanagari, i.e. a back-transliteration was performed. Words with language tag ‘en’ were labeled with their standard spelling. Words with language tag ‘rest’ were kept as they are.

3. **Part-of-Speech Tagging:** The universal Part-of-speech tagset [9] was used to label the POS of each word as this tagset is applicable to both English and Konkani words, and it contained a level of coarseness that suited our goals. The following case-specific guidelines were also observed:
    1. Sub-lexical code-mixed words were annotated based on their context, since POS is a function of a word in a given context.
    2. Words embedded in a sentence of another language were tagged as per context of the matrix language, irrespective of the POS tag of the word in its original language.

## 4 Experiments and Results

Here, we repeated the experiments performed by [2] and added new part to system. The original system first tokenizes an utterance into words. Then, a language identification module classifies each word as Konkani, English or Rest. Based on the language assigned, the Normalisation module runs the Konkani or English normalisers. In this section, we explain the POS Tagging system used after the Normalisation system.

### 4.1 Part-of-Speech Tagging System

Understanding the Part-of-Speech POS tagging, which provides a basic level of syntactic analysis for a given word or sentence. It was modeled as a sequence labeling task using CRFs [10] and SVM following paper. The feature set comprised of:

1. **Basic Word Features:** Word based features such as affixes, context and the word itself.
2. **LANG:** Language label of the token, obtained from the Language Identification system. This can have the values - ‘en’, ‘kn’ or ‘rest’.
3. **NORM:** Lexical features extracted from the normalised form of the word. These include linguistic features such as bound and free morphemes, suffixes, prefixes.
4. **TPOS:** Output of Twitter POS tagger [8] for the given word.
5. **KPOS:** Output of Konkani POS tagger[1] for the given word.

To obtain the Konkani POS tag output, the output from the normalisation module was used by transliterating Romanised Konkani words into WX-notations. Konkani POS tags were obtained using the Cdac Konkani POS tagger http://kbcs.in/tools.html. This POS Tagger is trained on WX-notation, thus English and ‘Rest’ words were transliterated to WX-notation. These transliterations along with the Konkani normalised data was sent to the POS tagger and final POS tag was obtained. The features ablation for the POS Tagger are shown in Sect. 4.1. Each feature was added only if it showed a positive increase in the system accuracy. Table 1 presents the obtained results.

[1] http://kbcs.in/tools.html.

**Table 1.** Token level POS Tagger Accuracy

| Features | CRF Accuracy (%) | SVM Accuracy (%) |
|---|---|---|
| BASELINE | 85.53 | 85.73 |
| +LANG | 86.53 | 87.53 |
| +NORM | 87.72 | 88.53 |
| +TPOS | 90.59 | 91.53 |
| +KPOS | 91.47 | 92.53 |

## 5 Conclusion and Future Work

In this Paper, we have focused on building first step of shallow parser for Konkani-English code-mixed social data. Through this paper we present our efforts at attempting various statistical methods for POS tagging of code-mixed social media data. We have attempted to build Part-of-speech tagger for this language pair, which we hope would result in better data-mining and sentiment analysis across the Indian subcontinent. We also create a standard dataset of 5088 code-mixed Konkani-English sentences for building supervised models of shallow parsing on this data which we consider as our immediate future work. In the future, we intend to continue creating more annotated code-mixed social media data. We intend to use this dataset to build tools for code-mixed data like morph analysers, chunkers and parsers.

## References

1. Hidayat, T.: An analysis of code switching used by Facebookers (a case study in a social network site). Sekolah Tinggi Keguruan dan Ilmu Pendidikan (STKIP) Siliwangi Bandung (2012)
2. Phadte, A., Thakkar, G.: Towards normalising Konkani-English code-mixed social media text. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pp. 85–94 (2017)
3. Vyas, Y., Gella, S., Sharma, J., Bali, K., Choudhury, M.: Pos tagging of English-Hindi code-mixed social media content. In: EMNLP, vol. 14, pp. 974–979 (2014)
4. Dey, A. Fung, P.: A Hindi-English code-switching corpus. In: LREC, pp. 2410–2413 (2014)
5. Bali, K., Vyas, Y., Sharma, J., Choudhury, M.: 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, EMNLP 2014 (2014)
6. Rao, D., Yarowsky, D.: Part of speech tagging and shallow parsing of Indian languages. Shallow Parsing for South Asian Languages (2007)
7. Jamatia, A., Gambäck, B., Das, A.: Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In: RANLP, pp. 239–248 (2015)
8. Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A.: Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics (2013)

9. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset (2011). arXiv preprint. arXiv:1104.2086
10. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML, vol. 1, pp. 282–289 (2001)
11. Sharma, A., Gupta, S., Motlani, R., Bansal, P., Srivastava, M., Mamidi, R., Sharma, D.M.: Shallow parsing pipeline for hindi-english code-mixed social media text (2016)