Word Level Language Identification in English Telugu Code Mixed Data

Sunil Gundapu
Language Technologies Research Centre
KCIS, IIIT Hyderabad
Telangana, India
gundapusunil@gmail.com

Radhika Mamidi
Language Technologies Research Centre
KCIS, IIIT Hyderabad
Telangana, India
radhika.mamidi@iiit.ac.in

Abstract

In a multilingual or sociolingual configuration Intra-sentential Code Switching (ICS) or Code Mixing (CM) is frequently observed nowadays. In the world most of the people know more than one language. The CM usage is especially apparent in social media platforms. Moreover, ICS is particularly significant in the context of technology, health and law where conveying the upcoming developments are difficult in ones native language. In applications like dialog systems, machine translation, semantic parsing, shallow parsing, etc. CM and Code Switching pose serious challenges. To do any further advancement in code-mixed data, the necessary step is Language Identification. So, in this paper we present a study of various models - Nave Bayes Classifier, Random Forest Classifier, Conditional Random Field (CRF) and Hidden Markov Model (HMM) for Language Identification in English - Telugu Code Mixed Data. Considering the paucity of resources in code mixed languages, we proposed CRF model and HMM model for word level language identification. Our best performing system is CRF-based with an f1-score of 0.91.

1 Introduction

Code switching is characterized as the use of two or more languages, diversity and verbal style by a speaker within a statement, pronouncement or discourse, or between different speakers or situations. This code switching can be classified as intersentential (the language switching is done at sentence boundaries) and intra-sentential (the alternation in a single discourse between two languages, where the switching occurs within a sentence). Code mixing is inconsistently elucidated in disparate subfields of linguistics and frequent examination of phrases, words, inflectional, derivational morphology and syntax use of a term as an equivalent to Code Mixing.

Code mixing is defined as the embedding of linguistic units of one language into utterance of another language. CM is not only used in commonly used spoken form of multilingual setting, but also used in social media websites in the form of comments and replies, posts and especially in chat conversations. Most of the chat conversations are in a formal or semi-formal setting and CM is often used. It commonly takes place in scenarios where a standard formal education is received in a language different than the persons native language or mother tongue.

Most of the case studies state that CM is very popular in the world in the present day, especially in countries like India, with more than 20 official languages. One such language, Telugu is a Dravidian language used by a total of 7.19% people of Telangana and Andra Pradesh states in India. Because of influence of English, most people use a combination of Telugu and English in conversations. There is lot of research work being carried out in code mixed data in Telugu as well as other Indian languages.

Consider this example sentence from a Telugu website[1] which consists of movie reviews, short stories and articles in cross script code-mixed languages. This example sentence illustrates the mix-

[1] www.chaibisket.com

ing being addressed in this paper:

Example: "John/**NE** nuvvu/**TE** exams/**EN** baaga/**TE** prepare/**EN** aithene/**TE** ,/**UNIV** first/**EN** classlo/**TE** pass/**EN** avuthav/**TE** ./**UNIV**". (Translation into English: John, you will pass in the first class, even if the exams are well prepared). The words followed by /NE, /TE, /EN and /UNIV correspond to Named Entity, Telugu, English and Universal tags respectively. In the above example some words exhibit morpheme level code mixing, like in **"classlo"** : **"class"** (English word) + **"lo"** (plural morpheme in Telugu). We also consider the clitiques like supere: super (English root word) + e (clitique) as code mixed.

We present some approaches to solve the problem of word level language identification in Telugu English Code Mixed data. The rest of this paper is divided into five sections.

In section 2, we discuss the related work, followed by data set and its annotation in section 3. Section 4 describes the approaches for language identification. And finally section 5 reports Results, Conclusion and Future Work.

2 Related Work

Research on code switching is decades old Gold (1967) and a lot of progress is yet to be made. Braj B., Kachru (1976) explained about the structure of multilingual languages organization and language dependency in linguistic convergence of code mixing in an Indian Perspective. The examination of syntactic properties and sociolinguistic constraints for bilingual code switching data was explored by Sridhar et al. (1980) who contended that how ICS shows impact on bilingual processing. According to Noor Al-Qaysi, Mostafa Al-Emran (2017) code switching in social networking websites like Facebook, Twitter and WhatsApp is very high. The result of such case studies indicated that 86.40% of the students using code switching on social networks, whereas 81% of the educators do so.

One of the basic tasks in text processing is Part of Speech (POS) Tagging and it is primary step in most of NLP. Word level Language Identification can be looked at as a task similar to POS tagging. For POS tagging task Nisheeth Joshi et al. (2013) developed a Hidden Markov Model based tagger in which they take one word in a sentence as a data point. To figure out this sequence labeling classification problem they consider the words as observations and POS tags as the hidden states. Kovida at el. (2018), with the use of internal structure and context of word perform POS tagging for Telugu-English Code Mixed data.

Language Identification issue has been addressed for English and Hindi CM data. For this problem Harsh at el. (2014) constructed a new model with the help of character level n-grams of a word and POS tag of adjoining words. Amitava Das and Bjrn Gambck (2014) modeled various approaches like unsupervised dictionary based approach and supervised SVM word level classification by considering contextual clues, lexical borrowings, and phonetic typing for LI in code mixed Indian social media text.

Ben King and Steven Abney (2013) elaborate the LI problem as a monolingual text of sequence labeling problem. The authors used a weakly supervised method for recognizing the language of words in the mixed language documents. For language identification in Nepali - English and Spanish - English data Utsab Barman et al. (2014) extract the features like word contextual clues, capitalization, char N-Grams and length of word and more features for each word. These features are input for the traditional supervised algorithms like K-NN and SVM.

Indian languages are relatively resource poor languages. Siva Reddy and Serge Sharoff (2011) are modeled a cross language POS Tagger for Indian languages. For an experimental purpose they used the Kannada language by using the Telugu resources because Kannada is more similar to Telugu. To do effective text analysis in Hindi English Code mixed social media data Sharma et al. (2016) are stimulate the shallow parsing pipeline. In this work the authors modeled a language identifier, a normalizer, a POS tagger and a shallow parser.

Most of the experiments have been carried out by using dictionaries, supervised classification models and Markov models for language identification. Finally, most of the authors concluded that language modeling techniques are more robust than dictionary based models.

We attempted this language identification problem with different classification algorithms to analyze the results. According to our knowledge, until now no work has been done on language identification in Telugu-English Code-mixed data.

3 Dataset Annotation and Statistics

We use the English Telugu code mixed dataset for language identification from the Twelfth International Conference on Natural Language Processing (ICON-2015). The dataset is comprised of Facebook posts, comments, replies to comments, tweets and WhatsApp chat conversations. This dataset contains 1987 code-mixed sentences. These sentences are tokenized into words. And the tokenized words of each sentence are separated by new line. The dataset contains 29503 tokens.

Language Label	Label Frequency	Percentage of Label
Telugu	8828	29.92
English	8886	30.11
Universal	11033	37.39
Named Entity	756	2.56

Table 1: Statistics of Corpus.

Each word is manually annotated with its POS tag and language identification tag which are: Telugu (TE), English (EN), Named Entity (NE) and Universal (UNIV). Out of the total data, 20% data is kept aside for testing and the remaining data is used for training the model.

4 Approaches for Word Level Language Identification (LI)

LI is the process of assigning a language identification label to each word in a sentence, based on both its syntax as well as its context. We have implemented baseline models using Nave Bayes Classifier and Random Forest Classifier with term frequency (TF) representation. We also performed experiments using Hidden Markov Model (HMM) with Viterbi Algorithm and Conditional Random Field (CRF) using python crfsuite as described in the following sections.

4.1 Nave Bayes Classifier using Count Vectors

The baseline Nave Bayes model implemented for language identification in English Telugu Code Mixed data produced estimable results. We are considered this LI problem similar to a Text Classification problem. We consider each sentence as a document and each word in it as a term, for which we calculate the conditional probability for each class label. The Nave Bayes assumption is that independence among the features, each feature is independent to the other features. We first convert the input code mixed words into TF(Term Frequency) - IDF(Inverse Document Frequency) vectors.

Word to Vector: Initially, the raw text data will be processed into feature vectors using the existing dataset. We used the TF-IDF Vector as the features for both the trigrams and character trigrams. TF-IDF calculates a score that represent the importance of a word in a document.

- TF: Number of times term word(w) appears in a document / Total no. of terms in the document
- Inverse Document Frequency (IDF): loge (Total no. of documents / No. of with term w in it)

We trained the model using set of **"n"** word vectors and corresponding labels.

- Input Data: Word vector of each word in corpus

$$(w_1\_vec, w_2\_vec, ..., w_n\_vec)$$

- Class Labels: (EN, TE, UNIV, NE)
- Model: **(Input word vectors)** → **(Labels)**

The baseline Nave Bayes model with character level TFIDF word vectors performed with an accuracy of 77.37% and 72.73% with trigram TFIDF word vectors.

4.2 Random Forest Classifier

Random forest is an ensemble classifier which selects the subset of random code mixed observations and subset of class labels or variables from training data. For each subset of sample of training observations it sets up a decision tree after extracting the features. All these decision trees are consolidated together to get the prediction using a voting procedure.

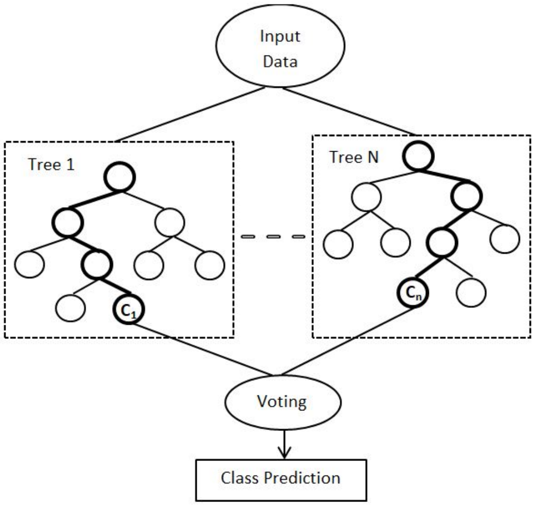

Figure 1: Random forest classifier model.

Initially, the whole observations in code mixed dataset are converted into vector form (Count Vector) for feature extraction. In a count matrix, a column represents a term from the dataset, a row represents the current word and the cell contains the frequency of the current word in the corpus.

We used 50 decision trees in the random forest for our experiments.

Random Forest automatically calculates the relevance score for each feature in training phase. The summation score of each feature is normalized to 1. These scores assist the model to select the significant features for model labeling. We use the **Mean Decrease in Impurity (MDI)** or **Gini Importance** to calculate the importance of each feature.

For this English Telugu Code Mixed corpus it performs with an accuracy of 77.34%, but since the dataset had very few sentences (1983 sentences, 31421 words) to construct the decision trees, we suspect overfitting was an issue.

4.3 Hidden Markov Model

HMM inherently takes sequence into account. We formulate the problem such that an entire sentence - sequence of words is a single data point. The observations are the words and the hidden states are language tags. This HMM based language tagging assigns the best tag to a word by calculating the forward and backward probabilities of tags along with the sequence provided as an input.

$$P(tag_i|word_i)$$

$$\Updownarrow$$

$$P(tag_i|tag_{i+1}) * P(tag_{i+1}|tag_i) * P(word_i|tag)$$

In above equation, $P(tag_i|tag_{i-1})$ probability for current tag given by the previous tag and $P(tag_{i+1}|tag_i)$ is the probability of the future tag given the current tag for focused word. Here the $P(tag_{i+1}|tag_i)$indirectly gives the meaning of transition of two tags.

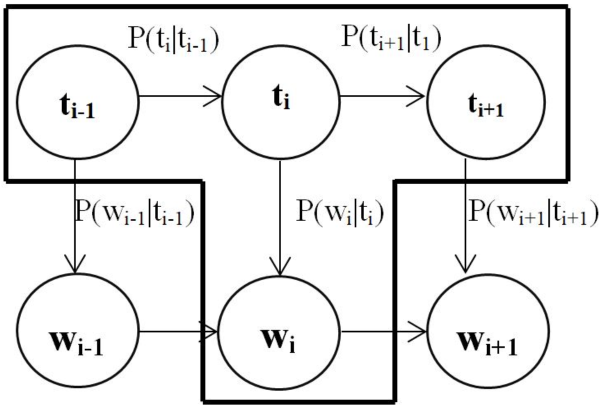

Figure 2: Context Dependency of HMM.

For each tag transition probability is computed by calculating the frequency count of two tags seen together in the corpus divided by the frequency count of the previous tag seen independently in the training corpus.

The likelihood probabilities calculated using $P(word_i|tag_i)$ i.e. the probability of the word given a current tag. This probability is computed using

$$P(word_i|tag_i)$$

$$\Updownarrow$$

$$Freq(tag_i, word_i)/Freq(tag_i)$$

Here, the probability of word provided a tag is computed by calculating the frequency count of the tag in the sentence and the word occurring together in the corpus divided by the frequency count of the occurrence of the tag alone in the corpus.

For testing the performance of the model, the corpus was divided into two parts: 80% for training, 20% for testing. The model performs with an accuracy of 85.46%. Perhaps, with more features, the accuracy could be further improved.

4.4 Conditional Random Field

Conditional Random Field (CRF) has been implemented for word - level language identification with the help of crf pysuite. CRF which is a simple, customizable and open source implementation. The most significant part of this approach is that feature selection. We proposed various feature set templates based on different possible combinations of available words, it's context and possible tags.

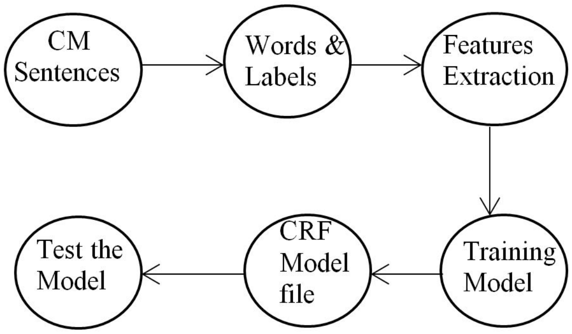

Figure 3: Steps involved in CRF Model.

For this model, we are use the following feature set to train the CRF Model.

- **Current word and its POS tag:** We consider the current testing word and its POS tag to predict the language label.
- **Next word and its POS tag:** We define the next word and its POS tag to capture the relation between current and next word.
- **Previous word and its POS Tag:** To extract the context of current word, we took the previous word and its POS tag based on first order markov assumption.
- **Prefix and Suffix of focus word:** Extracted the prefix and suffix of current word. If the word doesnt have any prefix or suffix we add NULL as a feature.
- **Length of word:** We considered the length of the word as one of the feature.
- **Start with numeric digit:** Whether the word starts with a numeric digit or not eg. **"2morrow"** (tomorrow).
- **Contains numeric digit:** Whether the word contains any number eg. **"ni8"** (night). A regular expression was used to detect this feature.
- **Start with special symbol:** Whether the word starts with any special symbol or character like !,@,$,/...etc.
- **Start with capital letter:** This feature is TRUE if the word starts with capital letter else FALSE.
- **Contains any capital letter:** Whether the word contains any capital letter like **"aLwAyS"** (always).
- **Previous word language tag:** To predict the current word language label we considered the previous word language label.
- **Character N-grams (Uni-, Bi-, Trigram of the word) :** Lot of code mixed words written in different formats. Example like (**akkada ekkada, yekkada, aeikkada**). To obtain the syntactical information of the word we took the uni-, bi-, trigram of the word in forward and backward direction into account. The uni-, bi-, trigrams of word in forward (**a, ak, akk**), backward (**a, da, ada**) of word respectively.

Above features are taken into account to tag the language label for current word. While testing, we assign untagged words such as URLs containing http:// or .in or www or smileys with a default tag of UNIV (universal tag).

Label	Precision	Recall	F1- Score
Telugu	0.84	0.81	0.87
English	0.88	0.87	0.88
NE	0.93	0.93	0.95
Universal	0.48	0.39	0.54
Average	0.89	0.90	0.91

Table 2: Experimental results of each tag.

With the help of scikit learn, Natural Language Took Kit and python crfsuite we performed the experiments on Engish Telugu corpus. The train corpus contains 23,635 words and test corpus contains 5868 words. We applied three-fold cross validation on our corpus for all experiments. The above feature set gave the highest accuracy of 91.2897%.

5 Results and Observations

The language identification was performed done by Naive Bayes and Random Forest classifiers as baseline models. Hidden Markov Model and CRF Model gave the best results for our problem. Comparatively, the HMM gave less accuracy than the CRF Model. The main reason for predicting the wrong language tag is the variation in tag used in the train data of English Telugu words. Our best performance system for tagging the language tag for a word is conditional random field with f1-score: 0.91 and accuracy: 91.2897%.

Model	Accuracy(%)
Nave Bayes Classifier	77.37
Random Forest Classifier	77.34
Hidden Markov Model	85.15
Conditional Random Field	91.28

Table 3: Consolidated Results (Accuracy).

In this work some interesting problems are encountered like Romanization of Telugu words, different types of syntax in social media text...etc. Since there is no standard way to transliterate the code mixed data and Romanization contributes a lot to the spelling errors in foreign words. For example, a single Telugu word can have the more than one spelling (Eg. “avaru”, “evaru”, “aivaru”, “yevaru”. Translation into English: “who”). This posed a significant challenge for language identification.

Similarly, In social media, chat conversation using SMS language “you” can be written as “U”, “Hai” “Hi”, “Good” “gooooood”...etc. Such non standard usage is an issue for language identification.

The results are encouraging and future work can be focused on obtaining more social media corpus and using deep learning approaches such as LSTM in future studies.

Acknowledgments

The authors would like to thank the Twelfth International Conference on Natural Language Processing (ICON -2015) for code-mixed dataset. We immensely grateful to Nikhilesh Bhatnagar for constructive criticism of the manuscript.

References

Sridhar, S. N., Sridhar, K. K. 1980. The syntax and psycholinguistics of bilingual code mixing. Canadian Journal of Psychology/Revue canadienne de psychologie, 34(4), 407-416.

Braj B. Kachru 1978. Toward Structuring Code-Mixing: An Indian Perspective. International Journal of the Sociology of Language, 16:27-46.

Noor Al-Qaysi, Mostafa Al-Emran. 2017. Code-switching Usage in Social Media: A Case Study from Oman. International Journal of Information Technology and Language Studies, 16:27-46.

Amitava Das, Anupam Jamatia, Bjrn Gambck. 2015. Part-of-Speech Tagging for Code-Mixed English Hindi Twitter and Facebook Chat Messages. . Proceedings of Recent Advances in Natural Language Processing.

Avinesh PVS, Karthik G. 2007. Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning . In Proceedings of Shallow Parsing for South Asian Languages.

Sharma, A., Gupta, S., Motlani, R., Bansal, P., Shrivastava, M., Mamidi, R., Sharma, D.M 2007. Shallow Parsing Pipeline - Hindi-English Code-Mixed Social Media Text . HLT-NAACL.

Iti M., Hemant D., Nisheeth J. 2013. HMM based pos tagger for Hindhi . International Conference on Computer Science and Information Technology.

Amitava Das, Bjrn Gambck. 2014. IdentifyingLanguages at the Word Level in CodeMixed Indian Social MediaText . Proceedings of the 11th International Conference on Natural Language Processing; 52.

Bali, K., Chittaranjan, G., Choudhury, M., Vyas, Y. 2014. Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System. CodeSwitch@EMNLP.

Jitta, D., Mamidi, R., Nelakuditi, K. 2016. Part of Speech Tagging for Code Mixed English-Telugu Social Media Data. Computational Linguistics and Intelligent Text Processing. CICLing 2016.

Reddy, S., Sharoff S. 2011. Cross Language POS Taggers (and other Tools) for Indian Languages : An Experiment with Kannada using Telugu Resources .

Computational Linguistics and the Information Need of Multilingual Societies.

Barman, U., Chrupala, G., Foster, J., Wagner, J. 2014. DCU-UVT: Word-Level Language Classification with Code-Mixed Data . CodeSwitch@EMNLP.

Abney, S., King, B. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. Association for Computational Linguistics.

Bhogi, S.K., Jhamtani, H., Raychoudhury, V. 2014. Word-level Language Identification in Bi-lingual Code-switched Texts . PACLIC.

Chandu K. Raghavi, Harsha P., Jitta D. Sai, Radhika M. 2017. Nee Intention enti? Towards Dialog Act Recognition in Code-Mixed Conversations. International Conference on Asian Language Processing (IALP-2017).

Kovida N. 2017. Towards Building a Shallow Parsing Pipeline for English-Telugu Code Mixed Social Media Data. Centre for Language Technologies Research Centre, IIIT/TH/2017/82.

Kumar, M.A., Soman, K.P., Veena, P.V. 2017. An effective way of word-level language identification for code-mixed facebook comments using word embedding via character-embedding. International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1552 - 1556.

Das, S.D., Das, D., Mandal, S. 2018. Language Identification of Bengali-English Code-Mixed data using Character and Phonetic based LSTM Models. CoRR, abs/1803.03859 .

Arnav S., Raveesh M. 2015. POS Tagging For Code Mixed Indian Social Media Text: Systems from IIITH for ICON NLP Tools Contest. Twelfth International Conference on Natural Language Processing (ICON-2015) .

Dong N., Seza Dogruoz A. 2013. Word level language identification in online multilingual communication. ACL 2013 .

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing.

Thamar, S., Elizabeth, B., Suraj, M., Steve, B., Mona, D., Mahmoud, G., Abdelati, H., Fahad, A., Julia, H., Alison, C., Pascale, F. 2014. Overview for the first shared task on language identification in code switched data.. In Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 62-72. Association for Computational Linguistics.