Identifying Languages at the Word Level in Assamese-Bengali-Hindi-English Code-Mixed Social Media Text

Neelakshi Sarma, Sanasam Ranbir Singh and Diganta Goswami
Indian Institute of Technology Guwahati
(s.neelakshi, ranbir, dgoswami)@iitg.ac.in

Abstract

Identifying languages at the word level is imperative for processing useful information from code-mixed social media text. Most existing studies in word level language identification are not suitable for low resource languages due to the unavailability of required resources like dictionaries, annotated resources, parsers, taggers etc. This paper aims to address the problem of word level language identification for low resource languages in a highly multilingual environment. Different word level language identification strategies are proposed and their performances are evaluated over a corpus of code-mixed transliterated text consisting of four languages - Assamese, Bengali, Hindi and English. From the experimental observations, it is evident that the proposed framework can effectively handle out-of-vocabulary and shared vocabulary problem which are major challenges for language identification over social media text in a multilingual environment. The proposed framework also uses minimal resources making it suitable for low resource languages.

Keywords

word language identification; code-mixing; social media text; assamese; bengali; hindi; english;

1. Introduction

Automatic Language Identification (ALI) is the task of automatically identifying languages in a given piece of text (Baldwin and Lui (2010, Los Angeles, CA)). With the explosion of multilingual content on the Web, identification of languages present in the text is an important pre-requisite for different applications like text-to-speech synthesis, machine translation, sentiment analysis etc. While ALI in the regular text domain (e.g. news articles, web documents etc.) is restricted to document or sentence level language identification, in the social media domain, this is often not enough. Phonetic typing (transliterated text) and code-mixing are a common phenomena in social media. Therefore, to extract important information from the text, it is necessary to identify languages at the word level. For example consider the text *Tumak sobe hate kore [Everybody hates you]* and *hate kame koribo lagibo [We have to work hands down]*. In the first case, *hate* is an English word embedded into a transliterated Assamese sentence while in the second case, *hate* is a transliterated Assamese word. For an application like sentiment analysis, it is important to identify that the word *hate* in the first case in an English word while the second is not. Therefore, in addition to the general chal-

lenges that social media text imposes on ALI owing to factors such as irregular and informal writings, mis-spellings, creative spellings etc., in a multilingual environment, code-mixing and the use of the same script to write content in different languages whether due to transliteration or due to shared script between languages imposes additional challenges to language identification.

Most existing literature in word level language identification suggest the use of different tools and resources for language identification like dictionaries (Barman et al. (2014, Doha, Qatar)), annotated resources (Barman et al. (2014, Doha, Qatar); Jaech et al. (2016)), monolingual corpora (King and Abney (2013)), transliteration tools (Singh et al. (2018)) etc. However, these approaches have a few short comings. First, the languages may actually lack the resources like dictionaries, annotated corpus, transliteration tools etc. Second, although the languages may consist some basic resources in the native script, due to the use of transliterated text for posting content in social media, these resources cannot be used. Third, with the amount of noise in the social media owing to mis-spellings, creative spellings, transliteration etc, traditional methods like dictionary look-up cannot be expected to achieve significantly good results. Fourth, resources like dictionaries or monolingual corpora are more suitable for independent word language classification. However, as evident in the previous examples, in a multilingual environment, the language of a word is context dependent.

Motivated by these challenges, this paper aims to address the problem of word level language identification for social media text in low resource languages in a highly multilingual environment. The work is motivated from the fact that obtaining sentence level annotations is far less expensive than obtaining word level annotations. Hence, more feasible for low resource languages. Therefore, this paper proposes a word level language identification framework using sentence level annotations. The proposed method focuses on learning word level representations by exploiting sentence level structural properties to build suitable word level language classifiers. Since social media text is noisy and evolving in nature, an annotated corpus is likely to miss out irregular word forms, new vocabulary or out-of vocabulary words. Therefore, the objective in this paper is to make use of the structural properties in the sentences to capture the language characteristics such that language labels can be obtained for new evolving words.

For our experimental study, we consider code-mixed corpora consisting of four languages - Assamese, Bengali, English and Hindi in transliterated text collected from a highly multilingual environment from two popular social media platforms - Facebook and YouTube. The paper presents and evaluates various word level language identification frameworks, and investigates their pros and cons. The experimental results reveal that the proposed framework can address out-of-vocabulary and shared vocabulary problem.

The rest of this paper is organized as follows. Related literature are discussed in Section 2. Proposed framework is discussed in Section 3. Experimental dataset is discussed in Section 4. Experimental set-ups and results are presented in Section 5. Conclusion and future works are discussed in Section 6.

2. Related Work

Language Identification is a well investigated problem. Most earlier works on language identification focused on language identification of well-edited documents like news-articles(Cavnar and Trenkle (1994)), web documents (Baldwin and Lui (2010, Los Angeles, CA)) etc. and have reported to have achieved very high accuracy. The character n-gram approach proposed by Cavnar and Trenkle (1994) achieved as high as 99.8% accuracy over a corpus of newsgroup articles. However, with the advent of social media platforms, language identification has become more challenging. Language identification of Twitter text at the tweet level

has been investigated in Carter et al. (2013) and Zubiaga et al. (2016). Carter et al. (2013) also show with systematic analysis that language identification of social media text is more challenging than that of formal text. However with the increasing amount of code-mixed multilingual data, language identification needs to accomplished at the word level for more efficient processing of information. This section presents a discussion on some of the existing studies on word level language identification.

Word level language identification in code-mixed text has been addressed in Barman et al. (2014, Doha, Qatar) where they experiment with different methods like dictionary look-up, supervised learning approach (SVM classifier) and sequence classification approach (CRF) over a set of transliterated Facebook comments. A similar study by the same authors have been reported in Das and Gamback (2014, Goa, India) where they perform an elaborate analysis on the nature of code-mixing in social media text. They also introduce a metric called Code-Mixing Index that indicate the level of mixing between different languages in a given text. Nguyen and Doğruöz (2013, Washington, USA) also explore different methods for word level language identification like dictionary look-up, language models, logistic regression classifier and conditional random fields classifier. Vyas et al. (2014) create a corpus of Facebook comments and also address back-transliteration, normalization and part-of-speech tagging in addition to language identification. CRF model using features like word features (word vector of the current word and that in the context), spelling features and intra-word features is used in Xia (2016, Texas, USA). CRF based system for word level language identification have also been used in Chittaranjan et al. (2014, Doha, Qatar). Features like modified edit distance, word frequency, character n-grams and part of speech tag and language tag of neighboring words for word language identification has been used in Jhamtani et al. (2014, Phuket, Thailand).

In contrast to the above studies that use code-mixed data collected from different social media platforms, Gella et al. (2014, Goa, India) create a synthetic code-mixed language identification dataset from a collection of monolingual text. They build binary classifiers for each language and then combine the results obtained from these classifiers to obtain the final language label for each word. Monolingual text has also been used in King and Abney (2013) where they use a collection of monolingual texts and employ weakly supervised and semi-supervised methods for word level language identification in multilingual documents. Rijhwani et al. (2017) also use an unsupervised model using monolingual corpora and HMM for word level language identification.

Neural network based approaches like Convolutional Neural Network and Bidirectional Long-Short-Term-Memory networks for word language identification have been explored in Jaech et al. (2016) for word level language identification. Recurrent Neural Networks have also been used in Samih et al. (2016) for word level language identification. Given a dataset of tweets annotated at the word level, they train their network for word level language predictions. Mandal and Singh (2018) use a multichannel neural network combining Convolutional Neural Networks and Long Short Term Memory Networks for word language identification in a corpus of code-mixed data. A sequence to sequence model for language identification have been explored in Jurgens et al. (2017). Language Identification is also addressed in Singh et al. (2018) where they use a transliteration model to transliterate romanized script to Devnagiri script. They use RNNs to train a character language model for each language. Given an annotated corpus, the output of the language models is combined with other features to train a three-class classifier which classifies a given word as belonging to either of the classes or not belonging to any of the classes.

Most of the works discussed above make use of different resources like dictionaries, annotated resources and transliteration tools which is not easy to obtain for low resource languages.

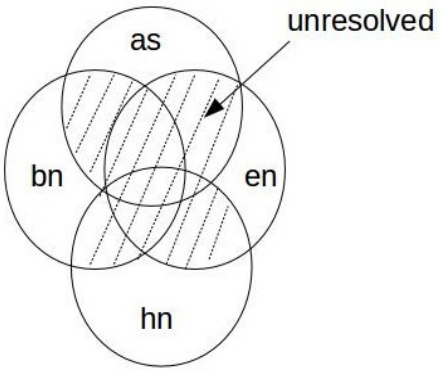


Figure 1: Obtaining word level annotations using sentence level annotations

A few studies like King and Abney (2013) and Rijhwani et al. (2017) do not use annotated resources. Instead they use monolingual corpus to obtain word level language labels in multilingual documents. However, this study concerns with word level language identification over a set transliterated text. Obtaining clean transliterated monolingual corpus is not possible because 1) transliteration in social media do not follow any fixed set of rules and 2) developing transliteration method for new languages is an additional overhead. Therefore, this study proposes to obtain word level language annotations using sentence level annotations. Since obtaining sentence level annotations is less expensive than obtaining word level annotations in terms of time and effort, the proposed method is clearly suitable for low resource languages.

3. Proposed Framework

This section presents the proposed framework to address word level language identification using sentence level annotations. The proposed method focuses on following key points ; (i) obtaining word level language information from the sentence level information (ii) language identification using global semantic similarity and (iii) language identification using local contextual similarity.

3.1. Obtaining word level language information from the sentence level information

Word level language identification in the traditional set-up make use of word level annotated data or monolingual corpora. While word level annotated data is expensive to obtain, it is also not possible to obtain clean monolingual transliterated text for low resource languages. A comparatively less expensive task is to obtain annotations at the sentence level. Therefore, the proposed model uses a dataset of sentences annotated with their corresponding languages to obtain word level annotations.

In the first step, the model uses the sentence level language information to group the words into two disjoint sets - *resolved set* and *unresolved set*. The resolved set of words are those words that occur only in sentences of one particular language and therefore these words can be assigned the language of the corresponding sentences. The *unresolved set* of words on the other hand is the set of words that occur in sentences of more than one language. Therefore, it is not clear at this stage whether the words belong to one particular language and have been borrowed or embedded in sentences of other languages or whether the words are valid in multiple languages. This is shown in Figure 1.

Let $\mathcal{S}$ be the set of annotated sentences, $\mathcal{V}$ be the set of all vocabulary in the annotated sentences and let $\mathcal{L}$ be the set of languages with which the sentences have been annotated.

Let l_v be the set of languages of the sentences in which the word v is occurring. Then if $|l_v| = 1$, i.e., the word v occurs in sentences of only one particular language, then we can safely assume that the word v belongs to language l. However, if $|l_v| > 1$, then the language of the word cannot be inferred at this stage. This word may be a legitimate word in more than one language. This word can also be a borrowed word belonging to another language. With $|l_v|$, $\mathcal{V}$ can be divided into two disjoint subsets ; (i) *resolved* set $\mathcal{R} = \{v : |l_v| = 1, v \in \mathcal{V}\}$ and (ii) *unresolved* set $\mathcal{U} = \{v : |l_v| > 1, v \in \mathcal{V}\}$ where $\mathcal{R} \cap \mathcal{U} = \phi$ and $\mathcal{R} \cup \mathcal{U} = \mathcal{V}$.

From manual verification over a selected set of words, we found that about 95% of the words in the resolved set $\mathcal{R}$ are correct. Considering this high accuracy, this paper focuses on resolving the words in the unresolved set $\mathcal{U}$.

3.2. Language identification using global semantic similarity

Word embeddings like word2vec (Mikolov et al. (2013)) or Glove (Pennington et al. (2014, Doha, Qatar)) is a well explored area for generating word representations from a given text corpus. They generate low dimensional vector representations of words that capture the syntactic and semantic characteristics of the words in the corpus by making use of the distributional characteristics of words within the corpus. Therefore the similarity or dis-similarity between two words can be determined by their vector representations. Due to their ability to capture latent relationships between the words, these representations have been found to be very useful in various text processing applications.

In this study, we make use of these representations to determine the language of a new word. The intuition is that since these representations capture the semantic similarity by using distributional characteristics in a text corpus, the language of a new word can be determined by measuring the similarity of its representation with representations of other words whose languages have already been determined. It is expected that any word will be semantically more similar to other words of the same language than words belonging to different language.

For obtaining the word representations, this study explores the word2vec word embedding method. A large set of user messages collected from Facebook and YouTube are used to obtain the word embeddings. Further, since the sentences used to obtain the word representations are code-mixed noisy sentences, the resulting word representations may also be noisy in nature i.e., words in different languages may lie closer compared to words in the same language in the projected vector space. We therefore further finetune the word representations using the triplet loss introduced in Schroff et al. (2015). The triplet loss ascertains that a sample x_i^a (anchor) belonging to a particular class is closer to all other samples x_i^p of the same class than to any sample x_i^n of another class. Therefore, the goal in triplet loss is to obtain the following:

$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2, \tag{1}$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \mathcal{T} \tag{2}$$

Here $f(x)$ denotes the embedding of word x and α denotes the margin that should be imposed between the positive pair and the negative pair. $\mathcal{T}$ is the set of all possible triplets (x_i^a, x_i^p, x_i^n). Then, if N is the cardinality of $\mathcal{T}$, then the loss to be minimized is given by

$$\mathcal{L} = \sum_{i}^{N} [||f(x_i^a - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha]_+ \tag{3}$$

We use the samples in the *resolved set* $\mathcal{R}$ as the training samples to generate the set of triplets to fine-tune the word2vec word embeddings using the loss discussed above.

The word representations discussed above are then used as features for training different classification frameworks. This study explores three different popularly used classification

frameworks - Support Vector Machines (SVM), Naive Bayes (NB) and Convolutional Neural Networks (CNN).

While building the classifier for word level language identification, we have considered the words in the resolved set $\mathcal{R}$ as the training samples. These classifiers are used to resolve the words in the unresolved set $\mathcal{U}$. Since $\mathcal{U} \cap \mathcal{R} = \phi$, none of the unresolved examples in $\mathcal{U}$ are present in the training set. This method may be suitable for capturing global language characteristics (identifying the borrowed words) but may not be suitable for capturing local characteristics (words that are valid in multiple languages). The method proposed in Section 3.3 attempt to capture the local characteristics.

3.3. Language Identification using local contextual similarity

An important issue that needs to be addressed while identifying languages of words is to disambiguate the language of words based on their context. This is particularly important when the languages in consideration are similar in terms of vocabulary used. Also, with transliterated text, the transliterated version of a word is often confused with valid words in a different language. This makes it important to consider the local context of the word in addition to the global semantics of the word. Let w be the target word that needs to be classified. Then, to classify the language of a word based on its local contextual similarity, a word is represented by a chunk of words that comprises of the target word w and the set of words on either side of w within a fixed window size h. The word embeddings discussed in in Section 3.2 are used as features and the words in the resolved set $\mathcal{R}$ are used as training samples to classify the words in the unresolved set $\mathcal{U}$.

4. Datasets

The objective of this paper is to study word level language identification for low resource languages in a highly multilingual environment. The study proposes to address this problem by exploiting structural properties of sentences and sentence level annotations. Though a few word level language identification frameworks are available online, they provide word level language information without providing the sentence level language information. Since such datasets do not fit into our proposed framework, we therefore locally generate two datasets from two popular social media platforms - Facebook and YouTube. While generating the datasets, we ensure the following criteria :

- Data should collected from a highly multilingual environment.
- There should be shared vocabulary across languages
- All the languages in the dataset should be in a common script (This study considers roman script)

This study considers Assam as the target location. Assam is a state in the North-Eastern part of India and majority of the population in Assam speak English, Hindi, Assamese, Bengali and a few ethnic languages. We identify various channels in Facebook and YouTube (Facebook groups and Facebook pages and YouTube channels) that are popular in the state of Assam. Data from the Facebook channels are collected using Facebook Graph API. Data from the YouTube channels are collected using the YouTube data API. A total of 409,168 and 15,829 user messages have been collected from Facebook and YouTube respectively. From this set, 28,968 number of Facebook messages and 11,486 number of YouTube messages are manually annotated with one of the four languages - Assamese, Bengali, Hindi and English. Details of the dataset are given in Table 1. While annotating the sentences, following guidelines have been considered.

- Messages in which words from different languages are mixed without changing the underlying language sense are considered as monolingual sentences and assigned the

	Facebook				YouTube			
Language	**Annotated Sentences**		**Annotated Words**		**Annotated Sentences**		**Annotated Words**	
	#messages	**Avg Length**	**Training**	**Testing**	**#messages**	**Avg Length**	**Training**	**Testing**
Assamese (as)	5,198	15 words	4,213	7,227	2,429	10 words	402	456
Bengali (bn)	1,594	12 words	1,036	2,355	129	8 words	247	252
Hindi (hn)	663	12 words	977	1,648	613	15 words	759	858
English (en)	21,531	24 words	22,805	22,020	8,315	12 words	1,366	1,336
Total	28,986	-	29,031	33,250	11,486	-	2,774	2,902

Table 1: Characteristics of the Experimental Datasets

Method	**Description**
M1	With word embedding of the current word as features, the words in the resolved set $\mathcal{R}$ are used as training samples
M2	With concatenation of the word embeddings of current word and the words in context as features, the words in the resolved set $\mathcal{R}$ are used as training samples
M3	With word embedding of the current word as features, training set consisting on manually annotated words
M4	With word embedding of current word and context word as features, training set consisting of manually annotated words

Table 2: List of experimental setups

language corresponding to the underlying language. For example *Ami edin success hom [We will be successful one day]* is an Assamese sentence with the English word success embedded in it. We assign this sentence as Assamese.

- Messages where two or more legitimate sentences in different languages are concatenated to form a conversational message are considered as multilingual messages. For example *Please dont mind. Ami amna e koisi. [Please dont mind. I just said like that.]*. Multilingual messages are currently ignored from this study.

To evaluate our methods, we have created an evaluation dataset annotated at the word level. Further, to compare the sentence level framework with that of the word level, we create another word level annotated dataset. This dataset is used to build the classifier with word level information. While annotating the words, the following guidelines have been considered.

- Words are annotated with one of the four languages according to the context. The same word may belong to different languages in different contexts. For example *Bad dau [Leave it]* and *Bad day*. The former is a phonetically type Bengali sentence while the later is an English sentence. The word *bad* in the first sentence is annotated as Bengali while that in the later sentence is annotated as English.
- Names entities do not belong to any language. They are therefore annotated with a different label - *ne*
- Universal expressions like *haha*, *hehe* etc. do not belong to any language and are annotated with a different label *amb*.

5. Experimental investigation and results

The experimental set-ups used in this paper are listed in Table 2. The same set of experimental set-ups are repeated with word2vec word embeddings and word2vec embeddings finetuned with triplet loss as discussed in section 3.2. With respect to the proposed framework, this

Language	Facebook		YouTube	
	#unique words	**#total words**	**#unique words**	**#total words**
Assamese (as)	12933	26588	5024	11256
Bengali (bn)	3183	5826	312	382
Hindi (hn)	1223	1679	1844	3014
English (en)	22548	98464	8390	26334
Unresolved	6336	495084	2398	97633

Table 3: Characteristics of Word labels obtained using sentence supervision

Using word2vec Word Embeddings												
	Facebook						**YouTube**					
	NB		**SVM**		**CNN**		**NB**		**SVM**		**CNN**	
	MacF	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**
M1	72.37	87.10	72.64	87.70	**72.77**	**87.32**	40.88	51.30	47.30	53.41	**73.79**	**80.19**
M2	82.64	87.01	84.56	**88.91**	**85.08**	88.63	54.06	60.57	54.10	58.99	**77.52**	**82.55**
M3	67.33	81.59	**74.45**	**88.76**	73.99	88.39	70.33	76.87	**75.52**	**83.01**	74.10	82.61
M4	82.55	86.94	82.55	90.23	**91.54**	**95.51**	74.35	79.08	51.93	57.78	**83.46**	**88.17**
Using Finetuned word2vec Word Embeddings												
	Facebook						**YouTube**					
	NB		**SVM**		**CNN**		**NB**		**SVM**		**CNN**	
	MacF	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**	**MacF**	**MicF**
M1	69.55	83.71	69.55	83.71	**72.35**	**86.77**	58.25	63.68	50.01	55.65	**71.58**	**79.62**
M2	84.18	88.49	85.83	90.13	**87.06**	**90.78**	67.58	73.81	59.25	63.05	**78.29**	**83.17**
M3	67.33	81.59	**74.45**	**88.76**	73.29	88.36	70.33	76.87	**75.52**	**83.01**	74.52	82.86
M4	83.89	87.88	91.86	95.36	**93.27**	**96.18**	76.97	81.52	54.61	59.54	**87.68**	**90.35**

Table 4: Macro Average F-scores (MacF) and Micro Average F-scores (MicF) for all experimental setups

study makes the following investigations :

- Confidence of the word level annotations obtained using the sentence level annotations
- Performance of word level language identification using global semantic similarity
- Performance of word level language identification using local contextual similarity
- Comparison of the performance of the proposed word level language identification to that of word level language identification using manually annotated word level data
- Language wise performance analysis
- Comparison between the original word2vec word representations and the word2vec word representations fine-tuned with the triplet loss
- Performance of the proposed framework in addressing out-of-vocabulary and shared vocabulary

5.1. Confidence of word level annotations obtained using the sentence level annotation

Section 3.1 describes the division of the words from the annotated sentences into resolved and unresolved set. Table 3 shows the number of words in different languages in the resolved set $\mathcal{R}$ and the unresolved set $\mathcal{U}$. It is important to evaluate the language classifications obtained at this stage for the resolved set because these words serve as the training samples for the subsequent stages. On comparison with the manually labeled data, the agreement on the language labels is 95.34%. Some of the cases where the annotations have gone wrong are : (i) words in

Using word2vec Word Embeddings									
		Facebook				**YouTube**			
		as	**bn**	**en**	**hn**	**as**	**bn**	**en**	**hn**
NB	**M1**	78.10	49.50	94.80	62.91	28.98	11.92	66.99	17.89
	M2	82.95	72.68	91.05	81.35	58.14	16.38	72.00	41.02
	M3	71.95	50.69	90.60	54.00	61.47	42.38	91.45	76.64
	M4	83.38	73.68	90.87	79.32	68.89	48.40	87.66	84.76
SVM	**M1**	79.84	49.01	95.30	60.87	31.25	17.39	68.13	21.11
	M2	84.77	77.55	92.54	81.02	51.87	13.86	70.11	81.22
	M3	79.73	48.77	96.38	63.97	71.46	56.46	91.81	81.45
	M4	91.55	86.30	97.28	85.45	27.00	30.72	69.14	42.97
CNN	**M1**	78.33	49.49	94.67	58.71	68.18	35.59	90.99	79.37
	M2	79.50	49.66	95.84	63.60	76.55	50.93	87.55	86.38
	M3	84.59	78.59	91.91	81.47	68.15	48.29	92.20	81.92
	M4	92.83	88.04	97.79	87.46	80.76	69.41	91.64	90.18
Using FineTuned word2vec Word Embeddings									
		Facebook				**YouTube**			
		as	**bn**	**en**	**hn**	**as**	**bn**	**en**	**hn**
NB	**M1**	74.96	47.90	92.54	60.98	54.47	40.55	75.38	42.56
	M2	85.90	76.29	91.91	80.04	72.86	44.35	80.93	68.68
	M3	71.95	50.69	90.60	54.00	61.47	76.64	91.45	76.64
	M4	85.46	78.24	91.16	77.42	72.61	54.02	88.04	87.30
SVM	**M1**	74.96	47.90	92.54	60.98	34.16	36.87	69.04	25.98
	M2	86.47	77.31	93.56	84.45	57.32	25.76	72.24	51.57
	M3	79.73	48.77	96.38	63.97	71.46	56.46	91.81	81.45
	M4	93.01	90.31	97.26	86.75	32.92	30.06	69.60	48.54
CNN	**M1**	78.35	50.34	94.87	61.20	68.06	47.90	89.42	80.78
	M2	88.69	80.27	93.78	85.67	74.69	53.08	89.47	84.86
	M3	78.89	51.06	95.63	64.75	68.74	57.64	92.85	8363
	M4	94.04	92.26	97.66	87.93	87.60	77.77	92.60	91.81

Table 5: Language wise F-score using different experimental set-ups

their noisy form have been borrowed into a sentence of another language but have not been used in their native language sentences e.g. words like handsme [handsome], vrfy [verify] occurring in sentences of languages other than English, (ii) words that have been misspelled when embedded in sentences of other language e.g. verivication [verification], reluxation [relaxation], (iii)low frequency words e.g. words like carnivorous, application which are legitimate English words but not occurring in English sentences but embedded in sentences of other languages.

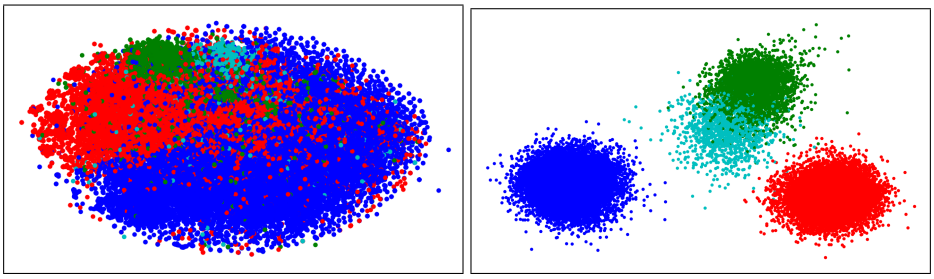
From these observations, it is expected that a larger dataset with a minimum support for all words will yield better results.

In this subsection, we have evaluated the performance of the word labels obtained directly using sentence labels. Since these annotations have been used to provide supervised information in the subsequent experiments, evaluations henceforth will only be based on the words that have been left unannotated at this stage i.e., the unannotated set $\mathcal{U}$ and will be annotated in the subsequent stages.

5.2. Performance of word level language identification using global semantic similarity
Word level language identification using global semantic similarity leads to independent word classification i.e., language classification of a word independent of the context. Thus, in this set-up, a word can belong to at-most one language. Using vector representations of words obtained using methods discussed in Section 3 as features, classifiers are trained using the words in $\mathcal{R}$. The words in $\mathcal{U}$ are then fed to the classifiers to obtain a language label for that word. Then given any sentence, the words in the sentence are labeled according to the label obtained using the classifier. However, the evaluation have been done considering the context in which the word has occurred. Results have been shown in table 4. It is seen that across all the classifiers, Convolutional Neural Networks show the best performance across both the datasets. A maximum macro-average F-score of 72.77% and 73.79% is achieved for the Facebook and YouTube datasets respectively. The not-so remarkable performance in this set-up can be attributed to the dependence of a word on the context. For example, the word *bad* has been classified as English. While it is correct in the English context e.g. *Bad day*, it is incorrect in the Assamese context e.g. *Bad dia [Leave it]* and in the Bengali context e.g. *bad dau [Leave it]*. Also, for the Facebook dataset, all the classifiers show similar performances while in the case of the YouTube dataset, there is considerable difference between the performance of CNN and the other classifiers. However, since the dataset for YouTube is much smaller, therefore it is expected that by increasing the size of the dataset, consistency in performance can be obtained.

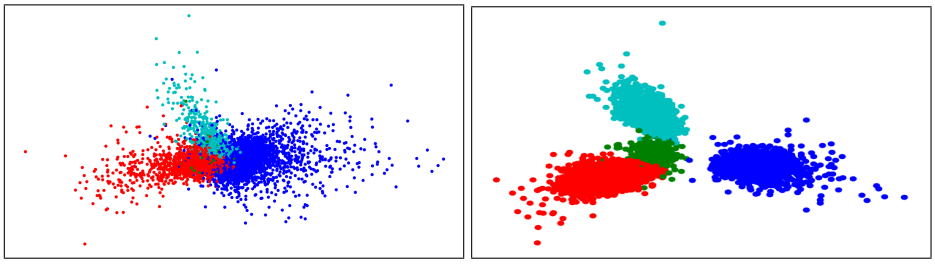
We further analyze the classification performance of the global similarity based method with regards to the embedded words i.e., words from a particular language that have been embedded or borrowed in sentences of other languages and the legitimate words i.e., words that are valid in multiple languages. It is observed that the percentage of embedded works correctly classified is 96.17% and 91.64% for the Facebook and YouTube datasets respectively and the percentage of legitimate words correctly classified is 76.59% and 65.60% respectively for the Facebook and YouTube datasets respectively. Thus it is evident that the independent word classification is capable of correctly identifying embedded words as compared to legitimate words. Thus if a corpus contains more number of embedded words than legitimate words, this method can be expected to perform better.

5.3. Word level language identification using local similarity
This set-up takes into account the fact that a word can belong to different languages depending on the context. Using the word vectors of the current word and the words in the context as features, classifiers are trained using the words in $\mathcal{R}$. The words within a window size of 3 on either side of the target word has been considered as features. The method M2 in table 2 refer to this set-up. It is seen that in this set-up as well, Convolutional Neural Networks show the best performance. A maximum macro-average F-score of 85.08% and 77.52% is achieved respectively for the Facebook and YouTube datasets using CNN classifier and word2vec word embeddings. This signifies an improvement of 16.91% and 5.05% over independent word classification (M1). Further, analyzing the performance with regard to the embedded words, i.e., words that belong to one particular language and have been borrowed or embedded in sentences of other languages and legitimate words i.e., words that are valid in more than one language, we observe that percentage of legitimate words correctly classified is 91.22% and 84.05% and the percentage of embedded words correctly classified is 86.98% and 83.12% for the Facebook and YouTube datasets respectively. While there is an improvement in the overall performance as well as in the performance of legitimate words correctly classified, there is a drop in the performance of classification of the embedded words as compared to that using global semantic similarity. However, an average F-score of 85.08% and 77.52%



(a) word2vec word embeddings (b) Fine-tuned word2vec word embeddings

Figure 2: 2-D visualization of word2vec word embeddings and finetuned word2vec word embeddings for the Facebook dataset



(a) word2vec word embeddings (b) Fine-tuned word2vec word embeddings

Figure 3: 2-D visualization of word2vec word embeddings and finetuned word2vec word embeddings for the YouTube dataset

without manually annotated word level corpus is an encouraging performance. Further, it is seen that the improvement in the YouTube dataset is considerably less compared to that of the Facebook dataset. However, since the YouTube dataset is very small in size compared to the Facebook dataset, it is expected that a larger dataset will further enhance the performance.

5.4. Sentence level vs word level

The classifiers described in the previous two setups have been trained using labels obtained using sentence level supervision. We repeat the same set-ups using a word level manually annotated dataset shown in Table 1. Referred to as M3 and M4 in Table 4, we see that though CNN and SVM show comparable performance, SVM slightly outperform CNN in the M3 set-up (global semantic similarity) for both the datasets while CNN shows the best performance in the M4 set-up (local contextual similarity). A maximum macro-average F-score of 74.45% and 91.54% is obtained using the global semantic similarity and local contextual similarity respectively for the Facebook dataset and 75.52% and 83.46% respectively for the YouTube dataset. It is observed that using global semantic similarity, the proposed framework using sentence level information obtain comparable performance performance to that using manually annotated word level corpus. Using local contextual similarity , the performance using sentence level information and word level annotated data are respectively 85.08% and 91.54% for the Facebook dataset and 77.52% and 83.46%. Clearly, training using word level annotated is giving better performance but considering that the proposed method is less expensive in terms of the required resources, we regard the performances as comparable.

An incoherent observation that is seen is that using SVM classifier, the global semantic

similarity based method (M3) shows better results than the local contextual similarity based method (M4) for the YouTube dataset. This is in contrast to all the remaining observations where local contextual similarity based method is giving better results than the global contextual similarity.

5.5. Analysis by language

Table 5 shows the language wise F-scores of the different classification methods. It is seen that in all cases, English (en) shows the best performance. If we look at the training set statistics, we see that the number of training samples for English far exceeds the other languages. This can be the reason of comparatively lesser performance for the other languages. Therefore, future efforts will be directed towards building a balanced training set.

5.6. Comparison between the original word2vec word representations and the word2vec word representations fine-tuned with the triplet loss

Table 4 shows that barring a few set-ups, using the fine-tuned word2vec word embeddings yields better performance than using the original word2vec embeddings in most set-ups. To examine the reason for the improved performance, we visualize the word embeddings for words in *resolved set* $\mathcal{R}$ in a two-dimensional space as shown in Fig 2 and Fig 3. Different colors denote different languages. We see that though words in the same language do form clusters in both the cases, there is considerable level of overlapping when using the original word2vec embeddings. On the other hand, using the fine-tuned word embeddings results in more distinct clusters. This explains the improved performance obtained using the finetuned word embeddings. However, it should be noted that the visualizations shown are only for the words in the resolved set $\mathcal{R}$ which are words that occur only in sentences of one particular language whereas the test set consists of words that occur across multiple languages. This explains why the improvement is not profound.

5.7. Performance of the proposed framework in addressing out-of vocabulary and shared vocabulary

The resolved set $\mathcal{R}$ and the unresolved set $\mathcal{U}$ are disjoint sets. Hence, none of the words in the test set are seen in the training set. Hence the macro-average F-scores of 87.06% (for the Facebook dataset using CNN classifier and fine-tuned word embeddings) and 78.29% (for the YouTube dataset using CNN classifier and fine-tuned word embeddings) show that the proposed method can effectively handle unseen or out-of-vocabulary words.

Further, all the words in the unresolved set $\mathcal{U}$ are words that occur in sentences of multiple languages i.e., they are shared across multiple languages. Therefore the performance obtained also show that the proposed framework is capable of addressing the shared vocabulary problem.

6. Conclusions and Future Work

This paper proposes different word level language identification techniques (using global semantic similarity and local contextual similarity) for low resource languages in a multilingual environment. The proposed methods make use of the sentence level structural properties and sentence level annotations to obtain word level annotations. Experimental observations show that performance comparable to that obtained using word level annotated data is obtained. The minimal resource requirements make the framework suitable for low resource languages. However, it is seen that while global semantic similarity is capable of identifying borrowed or embedded words, local contextual similarity is capable of resolving languages of words that are valid in multiple languages. Therefore, as a part of the future work, we would like

to explore methods that can combine the advantages of both the global semantic similarity and local contextual similarity. Further, the YouTube dataset used in this paper is very small. Therefore, a future work will be to build a larger and class balanced dataset.

References

Baldwin, T. and Lui, M., 2010, Los Angeles, CA, Language identification: The long and the short of the matter, in *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 229–237.

Barman, U., Das, A., Wagner, J. and Foster, J., 2014, Doha, Qatar, Code mixing: A challenge for language identification in the language of social media, in *Proceedings of The 1st Workshop on Computational Approaches to Code Switching*, pp. 13–23.

Carter, S., Weerkamp, W. and Tsagkias, M., 2013, Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text, *Language Resources and Evaluation Journal*, vol. 47, no. 1, pp. 195–215.

Cavnar, W. B. and Trenkle, J. M., 1994, N-gram-based text categorization, *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175.

Chittaranjan, G., Vyas, Y., Bali, K. and Choudhury, M., 2014, Doha, Qatar, Word-level language identification using crf: Code-switching shared task report of msr india system, in *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pp. 73–79.

Das, A. and Gamback, B., 2014, Goa, India, Identifying languages at the word level in code-mixed indian social media text, in *Proceedings of the 11th International Conference on Natural Language Processing*, pp. 378–387.

Gella, S., Bali, K. and Choudhury, M., 2014, Goa, India, "ye word kis lang ka hai bhai?" testing the limits of word level language identification, in *Proceedings of the 11th International Conference on Natural Language Processing*, pp. 368–377.

Jaech, A., Mulcaire, G., Hathi, S., Ostendorf, M. and Smith, N. A., 2016, Hierarchical character-word models for language identification, in *Conference on Empirical Methods in Natural Language Processing*, p. 84.

Jhamtani, H., Bhogi, S. K. and Raychoudhury, V., 2014, Phuket, Thailand, Word-level language identification in bi-lingual code-switched texts, in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pp. 348,357.

Jurgens, D., Tsvetkov, Y. and Jurafsky, D., 2017, Incorporating dialectal variability for socially equitable language identification, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 51–57, Association for Computational Linguistics, Vancouver, Canada, doi:10.18653/v1/P17-2009. URL `https://www.aclweb.org/anthology/P17-2009`.

King, B. and Abney, S., 2013, Labeling the languages of words in mixed-language documents using weakly supervised methods, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1110–1119.

Mandal, S. and Singh, A. K., 2018, Language identification in code-mixed data using multichannel neural networks and context capture, in *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pp. 116–120.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J., 2013, Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119.

Nguyen, D. and Doğruöz, A. S., 2013, Washington, USA, Word level language identification in online multilingual communication, in *Proceedings of the 2013 Conference on Empirical*

Methods in Natural Language Processing, pp. 857–862.

Pennington, J., Socher, R. and Manning, C., 2014, Doha, Qatar, Glove: Global vectors for word representation, in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Rijhwani, S., Sequiera, R., Choudhury, M., Bali, K. and Maddila, C. S., 2017, Estimating code-switching on twitter with a novel generalized word-level language detection technique, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1971–1982.

Samih, Y., Maharjan, S., Attia, M., Kallmeyer, L. and Solorio, T., 2016, Multilingual code-switching identification via lstm recurrent neural networks, in *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pp. 50–59.

Schroff, F., Kalenichenko, D. and Philbin, J., 2015, Facenet: A unified embedding for face recognition and clustering, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Singh, K., Sen, I. and Kumaraguru, P., 2018, Language identification and named entity recognition in hinglish code mixed tweets, in *Proceedings of ACL 2018, Student Research Workshop*, pp. 52–58.

Vyas, Y., Gella, S., Sharma, J., Bali, K. and Choudhury, M., 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)*, pp. 974–979.

Xia, M. X., 2016, Texas, USA, Codeswitching language identification using subword information enriched word vectors, in *Proceedings of The Second Workshop on Computational Approaches to Code Switching*, pp. 132–136.

Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A. and Fresno, V., 2016, Tweetlid: a benchmark for tweet language identification, *Language Resources and Evaluation*, vol. 50, no. 4, pp. 729–766.