Automatic Word-level Identification of Language in Assamese – English – Hindi Code-mixed Data

Manas Jyoti Bora, Ritesh Kumar
Department of Linguistics
Dr. Bhim Rao Ambedkar University, Agra
{manasjyotimj, riteshkrjnu}@gmail.com

Abstract

In this paper, we discuss the automatic identification of language in Assamese – English - Hindi code-mixed data at the word-level. The data for this study was collected from public Facebook Pages and was annotated using a minimal tagset for code-mixed data. Support Vector Machine was trained using the total tagged dataset of approximately 20k tokens. The best performing classifier achieved a state-of-the-art accuracy of over 96%.

Keywords: Code-mixing, Language Identification, Assamese, English, Hindi

1. Introduction

Code-mixing and code-switching in multilingual societies are two of the most well-studied phenomena within the field of sociolinguistics (Gumperz, 1964; Auer, 1984; Myers-Scotton, 1993; Muysken, 2000; Cardenas-Claros and Isharyanti, 2009 and several others). Generally, code-mixing is considered 'intra-sentential' in the sense that it refers to mixing of words, phrases or clauses within the same sentence while code-switching is 'inter-sentential' or even 'inter-clausal' in the sense that one switches to the other language while speaking. In this paper, we will use code-mixing to refer to both these phenomena.

While code-mixing is a very well-studied phenomena within the field of theoretical linguistics, there have been few works computational modelling of code-mixing. In the past of few years, with the explosion of social media and an urgent need to process the social media data, we have seen quite a few efforts at modelling, automatic identification and processing of code-mixing (most notable among them being Solorio and Liu, 2008a; Solorio and Liu, 2008b; Nguyen and Dogruoz, 2013; Das and Gambäck, 2014; Barman, et al., 2014; Vyas et al., 2014 and several others in the two workshops on computational approaches to code-mixing).

In this paper, we discuss the development of an automatic language identification system for Assamese-English-Hindi data at the word level. The data for this purpose was collected from Facebook pages. In the following sections, we discuss some of the previous works, with a focus on Indian languages, the method of corpus collection and annotation and the automatic language identification experiments.

Talking about the languages, English is quite well known and widely used language on online platforms in India. However, Assamese has recently become to be used pretty well in social media by the Assamese people. There was no prior work available on code-mixed social media content in Assamese when we began this particular work. Assamese and Hindi both are Indo-Aryan languages and therefore it is obvious to have many similarities between them, also due to contact and convergence. Specially the lexicons of the two languages have lot of word in common partly finding its root in Sanskrit and due to borrowing. From morphological perspective, we see that they are different in many ways. For instance, Assamese exhibits a rich inflectional morphological but also has agglutinating features in *classifiers* and *case markings*. In Hindi the Phi-features of *person, number* and *gender* are grammatical while in Assamese only the *person* is grammatical. Syntactically both the languages has the basic clause order of SOV.

Annotating the English words did not show much problem per se, however there is a lot of instances of misspelling. But while annotating Assamese and Hindi it was noticed that most of the time the spelling is not in standard form. There are many contractions and usage of non-canonical forms. Besides this there were instances where we saw that a single form was found among the languages which made it difficult to tag its language.

2. Previous Works in the Area

In the past few years, with growing interest and need for processing social media data, there have been quite few attempts at automatically recognising languages in code-mixed scenarios. While language identification at the document level across multiple languages (sufficiently different from each other) is generally considered a solved task, the same could not be claimed about code-mixed data. There have some attempts at language identification in Indian scenario, especially for Hindi-English (Vyas, et al. 2014; Das and Gambäck 2014), Bangla-English (Chanda, et al. 2016; Das and Gambäck 2014) and also Hindi-English-Bangla code-mixed data (Barman et al. 2014). These studies have shown that identifying language at the word-level, especially in the noisy, transliterated data of social media is a very significant and non-trivial task.

Das and Gambäck (2014) is one of the earliest works to address the problem of automatic identification of languages at word-level in social media Hindi-English as well as Bengali-English code-mixed data. They used a flat tagset with 17 tags with separate tags for named entity, abbreviation suffix in a different language. They use a simple dictionary-based method as the baseline and then go on to experiment with SVMs using 4 kinds of features – weighted character n-grams (3 and 4 grams were used), dictionary features (binary feature for each of the 3 languages, decided on the basis of presence / absence of the word in dictionary of a language), minimum edit distance weight (for out-of-dictionary words) and word context information (3 previous words with tags and 3 following words). The best performing system gave a high precision of over 90% (for Hindi-English texts) and 87% (for Bangla-English texts) but a low recall of 65% and 60% respectively, resulting in an overall F1 score of 76% and 74% respectively for Hindi and Bangla mixed texts. The performance of the system improved by 3% in case of Hindi and 2% in case of Bangla mixed texts

Vyas et al (2014) discuss the development of language identification system for Hindi-English mixed data in the context of developing a part-of-speech annotation system for social media data. They use a different kind of tagset that marks Matrix language of the sentence and Fragment language of the words. They used a word-level logistic regression (King and Abney 2013) for training their language identification system. The system was trained on 3201 English words from a SMS corpus and a separate Hindi corpus of 3218 words. The system gave an overall F1 score of 87% with a very low recall for Hindi data (since the data used for training did not contain spelling contractions and other variations and as such they were labelled as English by the classifier).

Chanda et al (2016), on the other hand, discusses the development of a system for identifying language in Bangla-English texts. They experiment with two different datasets – one from FIRE 2013 and the other of a Facebook Chat which they created. The best performing system makes use of Bangla and English dictionary (and the presence / absence of a word in the dictionary as a binary feature), n-gram and percentage of surrounding words that are predicted as Bangla (again using the dictionary). The model gives an F1 score of 91.5% for the FIRE dataset and 90.5% for the Facebook chat dataset, which is a big improvement over Das and Gambäck’s (2014) system but still not quite state-of-the-art.

The state-of-the-art system in identifying languages in code-mixed data in case of Indian languages is discussed by Barman et al (2014). Unlike the other studies, they experiment with a multilingual dataset and train their system on Hindi-English-Bangla code-mixed dataset. They use a tagset with 4 different tags – sentence, fragment, inclusion and wlcm (word-level code-mixing) – each with six attributes. A total of 2,335 posts and 9,813 comments, collected from a Facebook Group, were annotated with these tags. The best performing system was a CRF model trained using 5 different types of features – character n-grams, presence in dictionary, length of words, capitalization and contextual information (previous and next two words) – and it gave an accuracy of 95.76%, closely followed by an SVM model (trained with same features) with an accuracy of 95.52%.

As we could see, all of these approaches make use of language-specific dictionaries to train their models. One of our aims in this paper is to investigate if it is possible to build a reasonably good identification system without the use of a dictionary. Also till now there is no prior work on Assamese-English-Hindi code-mixed data and we plan to make some progress in that direction too.

3. Corpus Collection and Annotation

Since there is no previous corpus available for Assamese-Hindi-English, we collected a large corpus of such data from four different public Facebook pages:

- https://www.facebook.com/AAZGFC.Official
- https://www.facebook.com/Mr.Rajkumar007
- https://www.facebook.com/ZUBEENsOFFICIAL
- https://www.facebook.com/teenagersofassamm

The selection of the Facebook pages was not random. The users in these pages use code-mixing for various reasons. But first of all the users are from different sections of the society. There are different language users dominantly from Assamese who also use English in parallel. Hindi is used by a small number of users, besides Hindi is used in the pages mostly for funny comments with Assamese and also English together. There is one group of people who are seen to code-mix more than others – it is one of the reason for taking the particular pages. This kind of code-mixing has recently become popular among Facebook users. However this is not much common in speaking environments.

The first thing to start with after collecting the data is the annotation. This was done with the tool called ‘Webanno’ and the code-mixing tagset was used. Heavily inspired by Barman et al. (2014) and Vyas et al. (2014), the annotation scheme has three broad levels of annotation – matrix language (ML), fragment language (FL), and word-level code mixing (WLCM). Each of these levels could be annotated with one of the four language – Assamese (AS), English (EN), Hindi (HI) and Other (OT). The languages other than Assamese (AS) , English (EN) and Hindi (HI) are tagged as other (OT). The punctuations and sybmols are not marked separately and are given the same language name as the word preceding it. As defined earlier in previous works (Myers-Scotton, 1993), matrix language defines the grammatical structure of the sentence while the fragment language refers to the language whose words / phrases are mixed in a clause or a sentence. Word-level code mixing refers to the mixing at the word level – when the base morpheme is of one language and the bound morpheme (especially suffix) is of another language.

The annotation for this was carried out by a single annotator using Webanno. A total of 4768 comments with a total 20,781 tokens were annotated for the task. It took roughly a month to complete the annotation task. The detailed statistics is given in Table 1 below. A list of most frequent Hindi and English words mixed with Assamese (along with frequency of their mixing in the corpus) is also given Table 2.

Languages	**Token Count**
Assamese	11347
English	7689
Hindi	1200

Languages	**Token Count**
Others	545
Total	**20781**

Table 1: Token Count of each langage in the corpus

English	**Frequency**	**Hindi**	**Frequnecy**
u	147	और	19
you	114	hai	19
the	110	के	13
I	93	में	12
to	79	aap	12
of	76	ka	11
is	73	kya	10
day	73	इंडिया	9
love	62	को	8
a	61	हो	7

Table 2: Most frequent words mixed in Assamese

Let us also take a look at the data and where and why code-mixing occurs in the text.

Comment 1: 'khub enjoy karilu jua kali.'

(Facebook page: Zubeen Garg)

khub	enjoy	kar-il-u	jua kali
much	enjoy	do-PST-1	yesterday

"I/we enjoyed a lot yesterday (or last night)."

Even though there is a very common word for 'enjoy' in Assamese i.e. 'phurti', still 'enjoy' is used to express the feeling in a more profound way.

Comment 2:

“'जान्ना চাহ-তাহ হোঁ মে' [HI] মোৰ কথা হল, গানৰ মাজতো যে আপোনি "চাহ-তাহ" শ্বদটো লগাই অসমৰ চাহ পাত [AS] ইন্দাষ্টি্ৰলৈ (industry) [EN] যি বৰঙণি যোগালে, তাৰ কাৰনে মই অসম চৰকাৰক আপোনাৰ নামত এটা ৰাস্তা নাইবা [AS] এটলিস্ত (at least) [EN] এখন বাহঁৰ দলং, বা এখন কুকুৰা হাহঁ ছাগলিৰ আচনি হাতত লবলে আহ্বান জনাইছো [AS]”

(Facebook page: Mr. Rajkumar)

This comment is a ridicule because of the pronunciation of 'chahta' meaning 'want' as "চাহ-তাহ" which means 'tea~PRD' in a song. By this the commenter says that because the singer used the word 'tea', he has contributed to the Assam tea-industry.

Some of the other examples are given below.

Comment 3:

"Aji Sunday hoi toi gahori bonabi I know"

(Facebook page: Teenagers of Assam)

aji	sunday	hoi	toi	gahori	khabi	I know
today	sunday	be	you	pig	eat.FUT.2	I know

"Today is sunday so you will eat pork I know."

In this example a complete clause of English is mixed which is very commonly used in conversations among this group of speakers.

Comment 4:

"Moi 4 days continue apunar movie sai world record korim buli bhabisu"

(Facebook page: Teenagers of Assam)

	moi	4	days	continue	apunar	movie	sai
Gloss:	I	4	days	continue	your	movie	see.NF

	world	record	korim	buli	bhabisu
Gloss:	world	record	do.FUT	COMP	think.ASP.1

"I am thinking that I will make a world record by watching your movies for four days continuously."

4. Experiments

We experimented with Decision Trees and SVM for automatic classification of the language at the word-level (which is the 'fragment language' in our tagset). Our experiments included the following features:

Word Unigrams: This was the most basic feature (and equivalent to the use of dictionary in most of the previous studies).

Word Unigrams and Prefixes and Suffixes (upto 3): Character n-grams have generally proved to be very useful for the task of language identification. Also it has proved to be useful in similar tasks (Berman et al. 2014). Prefixes and suffixes are not actually character n-grams but we expect them to capture similar features of the text. The classifier trained using this feature set formed our baseline classifier.

Contextual Information: Different kinds of contextual information used for the experiments included tag of previous two words and previous and next two words. Again contextual information has proved to be very significant and useful in such tasks.

For the experiments, the data was split into 90:10 ratio with 90% used for training and 10% used for testing.

5. Results

As expected, SVM performed slightly better than the Decision Trees for this task and achieved an average accuracy of 96.01%. A comparative summary of the system's performance with different features is given in Table 3 below.

Features	Classifier	Precision	Recall	F1
Word	SVM	0.78	0.75	0.74
	DT	0.78	0.75	0.74
Word + All prefixes and Suffixes	SVM	0.81	0.81	0.80
	DT	0.80	0.79	0.79
Previous tag		0.93	0.93	0.93
		0.93	0.93	0.93
Previous Tag + Word	SVM	0.95	0.95	0.95
	DT	0.94	0.94	0.94
Previous 2 tags + Word + First Character	SVM	0.95	0.95	0.95
	DT	0.95	0.95	0.95
Previous 2 tags + Previous and next 2 words + word + 3 prefixes and suffixes	SVM	0.96	0.96	0.96
	DT	0.95	0.95	0.95

Table 3: Comparative scores of different feature combinations and classifiers

As could be seen from the above table, tag of the previous word plays probably the most important role in predicting the label of next work. The role of previous tag in such tasks have always been known to be significant and so it was expected that previous tag will play a significant role in the performance of the system. But what was a little surprising is the extent to which it affected the results. In fact, an F1 score of 93% is achieved just by using previous tag as the feature, which is much higher than any other combination of feature. Using words with the previous tags give a further 2% jump. And finally using the prefixes, suffixes and previous and next words, along with previous 2 tags and the word itself leads to a further 1% increase in the performance of the system. As is evident, our approach is language independent and it should work with any other language in a similar way. It pushes the current state-of-the-art by 0.25% and it might be possible to push it further with more data. It must be noted that Hindi is rather underrepresented in the dataset in comparison to English and Assamese. And a preliminary error analysis shows that the classifier achieves a very low precision of 77% with Hindi data. This aspect could definitely be improved with more data. Also further experiments with a sequence labelling algorithm like CRF might improve the results even further.

6. Summing Up

In this paper, we have discussed the creation of the first Assamese-Hindi-English code-mixed corpus, collected from Facebook and manually annotated. This corpus is being made available for further research. We also discussed the development of an automatic language identification system for code-mixed language. Our approach is language independent and could be used for developing similar systems for other languages also. In its current stage, the system gives an accuracy of 96.01%, which is 0.25% higher than the current state-of-the-art. We plan to carry out further experiments and hope to push the performance further up with different algorithms, making changes to the feature set and also by using more data (especially for Hindi) for training.

7. References

Arunavha, C., Das, D. and Mazumdar, C. (2016). Unraveling the English-Bengali Code-Mixing Phenomenon. In Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 80 – 89.

Auer, P. (1995). The pragmatics of code-switching: A sequential approach. In L. Milroy & P. Muysken (Eds.), *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching*. Cambridge: Cambridge University Press, pp. 115-135.

Barman, U., Das, A., Wagner, J. and Foster, J. (2014). Code Mixing: A Challenge for Language Identification in the Language of Social Media. In Proceedings of the First Workshop on Computational Approaches to Code Switching.

Cardenas-Claros, M. and Isharyanti, N. (2009). Code-switching and code-mixing in internet chatting: Between yes, ya, and si a case study. In The jaltcalljournal Vol. 5, No. 3 Pages 67–78.

Danet, B. and Herring, S. (2007). The Multilingual Internet: Language, Culture, and Communication Online. New York: Oxford University Press.

Das, A. and Gambäck, B. (2014). Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text. In Proceedings of the 11th International Conference on Natural Language Processing.

Eliasson, S. (1995). Myers-Scotton Carol, Duelling Languages: Grammatical Structure in Code-Switching. In *Language in Society,* Oxfored: Clarendon.

Gumperz, J. John (1964). Hindi-Punjabi Code-Switching in Delhi. In Proceedings of the Ninth International Congress of Linguistics. Mouton: The Hague.

King, B and Abney, S. (2013). Labeling the Languages of Words in Mixed-Language Documents Using Weakly

Supervised Methods. In Proceedings of NAACL-HLT, pages 1110–1119.

Muysken, P. (2000). Bilingual Speech: A Typology of Code-Mixing. Cambridge University Press.

Nguyen, D and Dogruoz, A. S. (2013). Word level Language Identification in Online Multilingual Communication. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 857–862.

Solorio, T. and Liu, Y. (2008a). Learning to Predict Code-Switching Points. In Proceedings of the Empirical Methods in Natural Language Processing

Solorio, T. and Liu, Y. (2008b). Parts-of-Speech Tagging for English-Spanish Code-Switched Text. In Proceedings of the Empirical Methods in Natural Language Processing.

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), pages 974–979, Doha, Qatar, October. Association for Computational Linguistics.