Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System

Gokul Chittaranjan
Microsoft Research India
t-gochit@microsoft.com

Yogarshi Vyas *
University of Maryland
yogarshi@cs.umd.edu

Kalika Bali **Monojit Choudhury**
Microsoft Research India
{kalikab, monojitc}@microsoft.com

Abstract

We describe a CRF based system for word-level language identification of code-mixed text. Our method uses lexical, contextual, character n-gram, and special character features, and therefore, can easily be replicated across languages. Its performance is benchmarked against the test sets provided by the shared task on code-mixing (Solorio et al., 2014) for four language pairs, namely, English-Spanish (En-Es), English-Nepali (En-Ne), English-Mandarin (En-Cn), and Standard Arabic-Arabic (Ar-Ar) Dialects. The experimental results show a consistent performance across the language pairs.

1 Introduction

Code-mixing and code-switching in conversations has been an extensively studied topic for several years; it has been analyzed from structural, psycholinguistic, and sociolinguistic perspectives (Muysken, 2001; Poplack, 2004; Senaratne, 2009; Boztepe, 2005). Although bilingualism is very common in many countries, it has seldom been studied in detail in computer-mediated-communication, and more particularly in social media. A large portion of related work (Androutsopoulos, 2013; Paolillo, 2011; Dabrowska, 2013; Halim and Maros, 2014), does not explicitly deal with computational modeling of this phenomena. Therefore, identifying code-mixing in social media conversations and the web is a very relevant topic today. It has garnered interest recently, in the context of basic NLP tasks (Solorio and Liu, 2008b; Solorio and Liu, 2008a), IR (Roy et al., 2013) and social media analysis (Lignos and Marcus, 2013). It should also be noted that the identification of languages due to code-switching is different from identifying multiple languages in documents (Nguyen and Dogruz, 2013), as the different languages contained in a single document might not necessarily be due to instances of code switching.

In this paper, we present a system built with off-the-shelf tools that utilize several character and word-level features to solve the EMNLP Code-Switching shared task (Solorio et al., 2014) of labeling a sequence of words with six tags viz. *lang1*, *lang2*, *mixed*, *ne*, *ambiguous*, and *others*. Here, *lang1* and *lang2* refer to the two languages that are mixed in the text, which could be English-Spanish, English-Nepali, English-Mandarin or Standard Arabic-dialectal Arabic. *mixed* refers to tokens with morphemes from both, *lang1* and *lang2*, *ne* are named entities, a word whose label cannot be determined with certainty in the given context is labeled *ambiguous*, and everything else is tagged *other* (Smileys, punctuations, etc.).

The report is organized as follows. In Sec. 2, we present an overview of the system and detail out the features. Sec. 3 describes the training experiments to fine tune the system. The shared task results on test data provided by the organizers is reported and discussed in Sec. 4. In Sec. 5 we conclude with some pointers to future work.

2 System overview

The task can be viewed as a sequence labeling problem, where, like POS tagging, each token in a sentence needs to be labeled with one of the 6 tags. Conditional Random Fields (CRF) are a reasonable choice for such sequence labeling tasks (Lafferty et al., 2001); previous work (King and Abney, 2013) has shown that it provides good performance for the language identification task as well. Therefore, in our work, we explored various token level and contextual features to build an optimal CRF using the provided training data. The features

* The author contributed to this work during his internship at Microsoft Research India

Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 73–79,
October 25, 2014, Doha, Qatar.

Lang.	Given Ids		Available		Available (%)	
	Train	Test	Train	Test	Train	Test
Es	11,400	3,014	11,400	1,672	100%	54.5%
Ne	9,999	3,018	9,296	2,874	93%	95.2%
Cn	999	316	995	313	99.6%	99.1%
Ar	5,839	2,363	5,839	2,363	100%	100%
Ar 2	-	1,777	-	1,777	-	100%

Table 2: Number of tweets retrieved for the various datasets.

Source	Language	For
instance_types_en.nt.bz2[1]	English	NE
instance_types_es.nt.bz2[1]	Spanish	NE
eng_wikipedia_2010_1M-text.tar.gz[2]	English	FW
spa_wikipedia_2011_1M-text.tar.gz[2]	Spanish	FW

Table 3: External resources used in the task. [1] http://wiki.dbpedia.org/Download, [2] http://corpora.uni-leipzig.de/download.html; NE:Named entities, FW:Word frequency list

used can be broadly grouped as described below:

Capitalization Features: They capture if letter(s) in a token has been capitalized or not. The reason for using this feature is that in several languages, capital Roman letters are used to denote proper nouns which could correspond to named entities. This feature is meaningful only for languages which make case distinction (e.g., Roman, Greek and Cyrillic scripts).

Contextual Features: They constitute the current and surrounding tokens and the length of the current token. Code-switching points are context sensitive and depend on various structural restrictions (Muysken, 2001; Poplack, 1980).

Special Character Features: They capture the existence of special characters and numbers in the token. Tweets contain various entities like hashtags, mentions, links, smileys, etc., which are signaled by #, @ and other special characters.

Lexicon Features: These features indicate the existence of a token in lexicons. Common words in a language and named entities can be curated into finite, manageable lexicons and were therefore used for cases where such data was available.

Character n-gram features: Following King and Abney (2013), we also used charagter n-grams for n=1 to 5. However, instead of directly using the n-grams as features in the CRF, we trained two binary *maximum entropy* classifiers to identify words of *lang1* and *lang2*. The classifiers returned the probability that a word is of *lang1* (or *lang2*), which were then binned into 10 equal buckets and used as features.

The features are listed in Table 1.

3 Experiments

3.1 Data extraction and pre-processing

The ruby script provided by the shared task organizers was used to retrieve tweets for each of the language pairs. Tweets that could not be downloaded either because they were deleted or protected were excluded from the training set. Table 2 shows the number of tweets that we were able to retrieve for the released datasets. Further, we found a few rare cases of tokenization errors, as evident from the occurrence of spaces within tokens. These were not removed from the training set and instead, the spaces in these tokens were replaced by an underscore.

3.2 Feature extraction and labeling

Named entities for English and Spanish were obtained from DBPedia instance types, namely, *Agent*, *Award*, *Device*, *Holiday*, *Language*, *MeansOfTransportation*, *Name*, *PersonFunction*, *Place*, and *Work*. Frequency lists for these languages were obtained from the Leipzig Copora Collection(Quasthoff et al., 2006); words containing special characters and numbers were removed from the list. The files used are listed in table 3. The *character n-gram* classifiers were implemented using the MaxEnt classifier provided in MALLET (McCallum, 2002). The classifiers were trained on 6,000 positive examples randomly sampled from the training set and negative examples sampled from both, the training set and from word lists of multiple languages from (Quasthoff et al., 2006); the number of examples used for each of these classifiers is given in Table 4.

We used CRF++ (Kudo, 2014) for labeling the tweets. For all language pairs, CRF++ was run under its default settings.

3.3 Model selection

For each language pair, we experimented with various feature combinations using 3-fold cross validation on the released training sets. Table 5 reports the token-level labeling accuracies for the various models, based on which the optimal feature sets for each language pairs were chosen. These optimal features are reported in Table 1, and the corresponding performance for 3-fold cross validation in Table 5. The final runs submitted for the

ID	Feature Description	Type	Features used in the final submission (Optimal set)			
			En-Es	**En-Ne**	**En-Cn**	**Ar-Ar**
Capitalization Features						
CAP1	Is first letter capitalized?	True/False	✓	✓	✓	NA
CAP2	Is any character capitalized?	True/False	✓	✓	✓	NA
CAP3	Are all characters capitalized?	True/False	✓	✓	✓	NA
Contextual Features						
CON1	Current Token	String	✓	✓	✓	✓
CON2	Previous 3 and next 3 tokens	Array (Strings)	✓	✓	✓	
CON3	Word length	String	✓	✓	✓	✓
Special Character Features						
CHR0	Is English alphabet word?	True/False			✓	NA
CHR1	Contains @ in locations 2-end	True/False	✓	✓	✓	✓
CHR2	Contains # in locations 2-end	True/False	✓	✓	✓	✓
CHR3	Contains ' in locations 2-end	True/False	✓	✓	✓	✓
CHR4	Contains / in locations 2-end	True/False	✓	✓	✓	✓
CHR5	Contains number in locations 2-end	True/False	✓	✓	✓	✓
CHR6	Contains punctuation in locations 2-end	True/False	✓	✓	✓	✓
CHR7	Starts with @	True/False	✓	✓	✓	✓
CHR8	Starts with #	True/False	✓	✓	✓	✓
CHR9	Starts with '	True/False	✓	✓	✓	✓
CHR10	Starts with /	True/False	✓	✓	✓	✓
CHR11	Starts with number	True/False	✓	✓	✓	✓
CHR12	Starts with punctuation	True/False	✓	✓	✓	✓
CHR13	Token is a number?	True/False	✓	✓	✓	✓
CHR14	Token is a punctuation?	True/False	✓	✓	✓	✓
CHR15	Token contains a number?	True/False	✓	✓	✓	✓
Lexicon Features						
LEX1	In lang1 dictionary of most frequent words?	True/False	✓	✓	✓	NA
LEX2	In lang2 dictionary of most frequent words?	True/False		✓	NA	NA
LEX3	Is NE?	True/False	✓	✓	NA	NA
LEX4	Is Acronym	True/False	✓	✓	NA	NA
Character n-gram Features						
CNG0	Output of two MaxEnt classifiers that classify lang1 vs. others and lang2 vs. others. This gives 2 probability values binned into 10 bns, two from each classifier, for the two classes.	Array (binned probability)	✓	✓	NA	NA
CRF Feature Type			U	U	U	B

Table 1: A description of features used. NA refers to features that were either not applicable to the language pair or were not available. B/U implies that the CRF has/does not have access to the features of the previous token.

Classifier	Languages used (And # words)
English-Spanish Language Pair	
Spanish vs Others	[es (6000)], [en (4000), fr (500), hi (500), it (500), po (500)]
English vs Others	[en (6000)], [es (4000), fr (500), hi (500), it (500), po (500)]
English-Nepali Language Pair	
Nepali vs Others	[ne (6000)], [en (3500), fr (500), hi (500), it (500), po (500)]
English vs Others	[en (6000)], [ne (3500), fr (500), hi (500), it (500), po (500)]
Standard Arabic vs. Arabic Dialects	
Std vs. Dialect	[lang1 (9000)], [lang2 (3256)]

Table 4: Data to train *character n-gram* classifiers.

shared task, including those for the surprise test sets, use the corresponding optimal feature sets for each language pair.

Feature	Context	Language Pair				
		En-Es	En-Ne†	En-Cn	Ar-Ar	Ar-Ar (2)
Development Set						
All	B	92.8	94.3	93.1	85.5	-
- CON2	B	93.8	95.6	94.9	81.2	-
- CHR*	B	92.3	93.5	91.0	85.3	-
- CAP*	B	92.7	94.2	90.1	-	-
- CON2	U	93.0	94.3	93.1	85.6	-
- CNG0	B	92.7	94.2	-	-	-
- LEX*	B	92.7	94.1	-	-	-
Optimal	-	95.0	95.6	95.0	85.5	-
Results on Test data for the optimal feature sets						
Regular		85.0	95.2	90.4	90.1	53.6
Surprise		91.8	80.8	-	65.0	-

Table 5: The overall token labeling accuracies (in %) for all language pairs on the training and test datasets. "-" indicates the removal of the given feature. '*" is used to indicate a group of features. Refer tab. 1) for the feature Ids and the **optimal** set. *B* and *U* stand for bigram and unigram respectively, where the former refers to the case when the CRF had access to features of the current and previous tokens, and the latter to the case where the CRF had access only to the features of the current token. †: Lexical resources available for *En* only.

4 Results and Observations

4.1 Overall token labeling accuracy

The overall token labeling accuracies for the regular and surpise test sets (wherever applicable) and a second set of dialectal and standard Arabic are reported in the last two rows of Table 5. The same table also reports the results of the 3-fold cross validation on the training datasets. Several important observations can be made from these accuracy values.

Firstly, accuracies observed during the training phase was quite high ($\sim$ 95%) and exactly similar for En-Es, En-Ne and En-Cn data; but for Ar-Ar dataset our method could achieve only up to 85% accuracy. We believe that this is due to unavailability of any of the lexicon features, which in turn was because we did not have access to any lexicon for dialectal Arabic. While complete set of lexical features were not available for En-Cn as well, we did have English lexicon; also, we noticed that in the En-Cn dataset, almost always the En words were written in Roman script and the Cn words were written in the Chinese script. Hence, in this case, script itself is a very effective feature for classification, which has been indirectly modeled by the CHR0 feature. On the other hand, in the Ar-Ar datasets, both the dialects are written using the same script (Arabic). Further, we found that using the CNG0 feature that is obtained by training a character n-gram classifier for the language pairs resulted in the drop of performance. Since we are not familiar with arabic scripts, we are not sure how effective the character n-gram based features are in differentiating between the standard and the dialectal Arabic. Based on our experiment with CNG0, we hypothesize that the dialects may not show a drastic difference in their character n-gram distributions and therefore may not contribute to the performance of our system.

Secondly, we observe that effectiveness of the different feature sets vary across language pairs. Using all the features of the previous words (context = B) seems to hurt the performance, though just looking at the previous 3 and next 3 tokens was useful. On the other hand, in Ar-Ar the reverse has been observed. Apart from lexicons,

character n-grams seems to be a very useful feature in En-Es classification. As discussed above, CHR* features are effective for En-Cn because, among other things, one of these features also captures whether the word is in Roman script. For En-Ne, we do not see any particular feature or sets of features that strongly influence the classification.

The overall token labeling accuracy of the shared task runs, at least in some cases, differ quite significantly from our 3-fold cross validation results. On the regular test sets, the results for En-Ne is very similar to, and En-Cn and Ar-Ar are within expected range of the training set results. However, we observe a 10% drop in En-Es. We observe an even bigger drop in the accuracy of the second Ar-Ar test set. We will discuss the possible reason for this in the next subsection. The accuracies on the surprise sets do not show any specific trend. While for En-Es the accuracy is higher by 5% for the surprise set than the regular set, En-Ne and Ar-Ar show the reverse, and a more expected trend. The rather drastic drops in the accuracy for these two pairs on the surprise sets makes error analysis and comparative analysis of the training, test and surprise datasets imperative.

4.2 Error Analysis

Table 6 reports the F-scores for the six labels, i.e., *classes*, and also an overall tweet/post level accuracy. The latter is defined as the percentage of input units (which could be either a tweet or a post or just a sentence depending on the dataset) that are correctly identified as either code-mixed or monolingual; an input unit is considered code-mixed if there is at least one word labeled as *lang1* and one as *lang2*.

For all the language pairs other than Arabic, the F-score for NE is much lower than that for *lang1* and *lang2*. Thus, the performance of the system can be significantly improved by identifying NEs better. Currently, we have used lexicons for only English and Spanish. This information was not available for the other languages, namely, Nepali, Mandarin, and Arabic. The problem of NE detection is further compounded by the informal nature of sentences, because of which they may not always be capitalized or spelt properly. Better detection of NEs in code-mixed and informal text is an interesting research challenge that we plan to tackle in the future.

Note that the *ambiguous* and *mixed* classes can be ignored because their combined occurrence is less than 0.5% in all the datasets, and hence they have practically no effect on the final labeling accuracy. In fact, their rarity (especially in the training set) is also the reason behind the very poor F-scores for these classes. In En-Cn, we also observe a low F-score for *other*.

In the Ar-Ar training data as well as the test set, there are fewer words of *lang2*, i.e., dialectal Arabic. Since our system was trained primarily on the context and word features (and not lexicon or character n-grams), there was not enough examples in the training set for *lang2* to learn a reliable model for identifying *lang2*. Moreover, due to the distributional skew, the system learnt to label the tokens as *lang1* with very high probability. The high accuracy in the Ar-Ar original test set is because 81.5% of the tokens were indeed of type *lang1* in the test data while only 0.26% were labeled as *lang2*. This is also reflected by the fact that though the F-score for *lang2* in Ar-Ar test set is 0.158, the overall accuracy is still 90.1% because F-score for *lang1* is 94.2%.

As shown in Table 7, the distribution of the classes in the second Ar-Ar test set and the surprise set is much less skewed and thus, very different from that of the training and original test sets. In fact, words of *lang2* occur more frequently in these sets than those of *lang1*. This difference in class distributions, we believe, is the primary reason behind the poorer performance of the system on some of the Ar-Ar test sets.

We also observe a significant drop in accuracy for En-Ne surprise data, as compared to the accuracy on the regular En-Ne test and training data. We suspect that it could be either due to the difference in the class distribution or the genre/style of the two datasets, or both. An analysis of the surprise test set reveals that a good fraction of the data consist of long song titles or part of the lyrics of various Nepali songs. Many of these words were labeled as *lang2* (i.e., Nepali) by our system, but were actually labeled as NEs in the gold annotations[1] While song titles can certainly be considered as NEs, it is very difficult to identify them without appropriate resources. It should however be noted that the En-Ne surprise set has only 1087 tokens, which is too small to base any strong claims or conclusions on.

[1] Confirmed by the shared task organizers over email communication.

Language Pair	F-measure (Token-level)						Accuracy of
	Ambiguous	lang1	lang2	mixed	NE	Other	Comment/Post
En-Es	0.000	0.856	0.879	0.000	0.156	0.856	82.1
En-Ne	-	0.948	0.969	0.000	0.454	0.972	95.3
En-Cn	-	0.980	0.762	0.000	0.664	0.344	81.8
Ar-Ar	0.000	0.942	0.158	-	0.577	0.911	94.7
Ar-Ar (2)	0.015	0.587	0.505	0.000	0.424	0.438	71.4
En-Es Surprise	0.000	0.845	0.864	0.000	0.148	0.837	81.5
En-Ne Surprise	-	0.785	0.874	-	0.370	0.808	71.6
Ar-Ar Surprise	0.000	0.563	0.698	0.000	0.332	0.966	84.8

Table 6: Class-wise F-scores and comment/post level accuracy of the submitted runs.

Dataset	Percentage of					
	Amb.	lang1	lang2	mixed	NE	Other
Training	0.89	66.36	13.60	0.01	11.83	7.30
Test-1	0.02	81.54	0.26	0.00	10.97	7.21
Test-2	0.37	32.04	45.34	0.01	13.24	9.01
Surprise	0.91	22.36	57.67	0.03	9.13	9.90

Table 7: Distribution (in %) of the classes in the training and the three test sets for Ar-Ar.

5 Conclusion

In this paper, we have described a CRF based word labeling system for word-level language identification of code-mixed text. The system relies on annotated data for supervised training and also lexicons of the languages, if available. Character n-grams of the words were also used in a MaxEnt classifier to detect the language of a word. This feature has been found to be useful for some language pairs. Since none of the techniques or concepts used here is language specific, we believe that this approach is applicable for word labeling for code-mixed text between any two (or more) languages as long as annotated data is available.

This is demonstrated by the fact that the system performs more or less consistently with accuracies ranging from 80% - 95% across four language pairs (except for the case of Ar-Ar second test set and the surprise set which is due to stark distributional differences between the training and test sets). NE detection is one of the most challenging problems, improving which will definitely improve the overall performance of our system. It will be interesting to explore semi-supervised and unsupervised techniques for solving this task because creating annotated datasets is expensive and effort-intensive.

References

Jannis Androutsopoulos. 2013. Code-switching in computer-mediated communication. In *Pragmatics of Computer-mediated Communication*, pages 667–694. Berlin/Boston: de Gruyter Mouton.

Erman Boztepe. 2005. Issues in code-switching: competing theories and models. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 3.2.

Marta Dabrowska. 2013. Functions of code-switching in polish and hindi facebook users' post. *Studia Linguistica Universitatis Lagellonicae Cracoviensis*, 130:63–84.

Nur Syazwani Halim and Marlyana Maros. 2014. The functions of code-switching in facebook interactions. In *Proceedings of the International Conference on Knowledge-Innovation-Excellence: Synergy in Language Research and Practice; Social and Behavioural Sciences*, volume 118, pages 126–133.

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL-HLT*, pages 1110–1119.

Taku Kudo. 2014. Crf++: Yet another crf toolkit. `http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar#links`, Retrieved 11.09.2014.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 282–289.

Constantine Lignos and Mitch Marcus. 2013. Toward web-scale analysis of codeswitching. In *87th Annual Meeting of the Linguistic Society of America*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Pieter Muysken. 2001. The study of code-mixing. In *Bilingual Speech: A typology of Code-Mixing*. Cambridge University Press.

Dong Nguyen and A. Seza Dogruz. 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.

John C. Paolillo. 2011. Conversational codeswitching on usenet and internet relay chat. *Language@Internet*, 8.

Shana Poplack. 1980. Sometimes i'll start a sentence in Spanish y termino en espanol: Toward a typology of code-switching. *Linguistics*, 18:581–618.

Shana Poplack. 2004. Code-switching. In U. Ammon, N. Dittmar, K.K. Mattheier, and P. Turdgill, editors, *Soziolinguistik. An international handbook of the science of language*. Walter de Gruyter.

U. Quasthoff, M. Richter, and C. Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the fifth International Conference on Language Resource and Evaluation*, pages 1799–1802.

Rishiraj Saha Roy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. 2013. Overview and datasets of fire 2013 track on transliterated search. In *Proceedings of the FIRE 2013 Shared Task on Transliterated Search*.

Chamindi Dilkushi Senaratne, 2009. *Sinhala-English code-mixing in Sri Lanka: A sociolinguistic study*, chapter Code-mixing as a research topic. LOT Publications.

Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the Empirical Methods on Natural Language Processing (EMNLP)*, pages 973–981.

Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the Empirical Methods on Natural Language Processing (EMNLP)*, pages 1051–1060.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steve Bethard, Mona Diab, Mahmoud Gonheim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirshberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identifiation in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching. Conferencfe on Empirical Methods in Natural Language Processing*.