Part-of-Speech Tagging for Code-Mixed English-Hindi Twitter and Facebook Chat Messages

Anupam Jamatia
National Institute of Technology
Agartala, Tripura, India
anupamjamatia@gmail.com

Björn Gambäck
Norwegian University of Science and Technology
Trondheim, Norway
gamback@idi.ntnu.no

Amitava Das
Indian Institute of Information Technology
Sri City, Andhra Pradesh, India
amitava.das@iiits.in

Abstract

The paper reports work on collecting and annotating code-mixed English-Hindi social media text (Twitter and Facebook messages), and experiments on automatic tagging of these corpora, using both a coarse-grained and a fine-grained part-of-speech tag set. We compare the performance of a combination of language specific taggers to that of applying four machine learning algorithms to the task (Conditional Random Fields, Sequential Minimal Optimization, Naïve Bayes and Random Forests), using a range of different features based on word context and word-internal information.

1 Introduction

Code-mixing occurs when a person changes language (alternates or switches code) below clause level, so internally inside a sentence or an utterance. This phenomenon is more abundant in more informal settings — such as in conversational spoken language and in social media text — and of course also more common in areas of the world where people are naturally bi- or multilingual, that is, in regions where languages change over short geospatial distances and people generally have at least a basic knowledge of the neighbouring languages. In particular, India is home to several hundred languages, with language diversity and dialectal changes instigating frequent code-mixing.

We will here look at the tasks of collecting and annotating code-mixed English-Hindi social media text, and on automatic part-of-speech (POS) tagging of these code-mixed texts. In contrast, most research on part-of-speech tagging has so far concentrated on more formal language forms, and in particular either on completely monolingual text or on text where code alternation occurs above the clause level. Most research on social media text has, on the other hand, concentrated on English tweets, whereas the majority of these texts now are written in other media and in other languages — or in mixes of languages.

Today, code-switching is generally recognised as a natural part of bi- and multilingual language use, even though it historically often was considered a sub-standard use of language. Conversational spoken language code-switching has been a common research theme in psycho- and sociolinguists for half a century, and the first work on applying language processing methods to code-switched text was carried out in the early 1980s (Joshi, 1982), while code-switching in social media text started to be studied in the late 1990s (Paolillo, 1996). Still, code alternation in conventional texts is not so prevalent as to spur much interest by the computational linguistic research community, and it was only recently that it became a research topic in its own right, with a code-switching workshop at EMNLP 2014 (Solorio et al., 2014), and a shared tasks at EMNLP and at Forum for Information Retrieval Evaluation, FIRE 2014.

Both these shared tasks were on automatic word-level language detection in code-mixed text, but here we will assume that the word-level languages are known and concentrate on the task of automatic part-of-speech tagging for these types of texts. We have collected a corpus consisting of Facebook messages and tweets (which includes all

Proceedings of Recent Advances in Natural Language Processing, pages 239–248,
Hissar, Bulgaria, Sep 7–9 2015.

possible types of code-mixing diversity: varying number of code alternation points, different syntactic mixing and language change orders, etc.), and carried out several experiments on this corpus to investigate the problem of assigning POS tags to code-mixed text.

The rest of the paper is organized as follows: In Section 2, we discuss the background and related work on part-of-speech tagging, social media text processing, and code-switching. The collection and annotation of a code-mixed corpus are described in Section 3, which also compares the complexity of the corpus to several other code-mixed corpora based on a code-mixing index. The actual part-of-speech tagging experiments are discussed in Section 4, starting by describing the features used, and then presenting the performance of four different machine learning methods. The results are elaborated on in Section 5, in particular how system performance is affected by the level of code-mixing, while Section 6 sums up the discussion and points to directions for future research.

2 Background and Related Work

In essence, this paper is concerned with the intersection of three topics: part-of-speech tagging, processing of social media text, and code-switching. In the present section, we will mainly discuss work related to the latter two topics, and tagging in relation to those.

First though, it should be noted that present-day POS taggers more or less receive 96+% performance on English news text with just about any method, with state-of-the-art systems going beyond the 97% point on the English Wall Street Journal corpus: Spoustová et al. (2009) report achieving an accuracy of 97.43% by combining rule-based and statistically induced taggers. However, most work on POS tagging has so far concentrated on a few European and East Asian languages, and on fairly formal texts, that is, texts of a quite different nature than the ones that are the topic of the present work.

2.1 Social Media and Code-Switching

The term 'social media text' will be used throughout this paper as referring to the way these texts are communicated, although it is important to keep in mind that social media in itself does not constitute a particular textual domain. Rather, there is a wide spectrum of different types of texts transmitted in this way, as discussed in detail by, e.g., Eisenstein (2013) and Androutsopoulos (2011). They both argue that the common denominator of social media text is not that it is 'noisy' and informal *per se*, but that it describes language in (rapid) change, which in turn has major implications for natural language processing: if we build a system that can handle a specific type of social media text today, it will be outdated tomorrow. Something which makes it very attractive to apply machine learning and adaptive techniques to the problem.

In all types of social media, the level of formality of the language depends more on the style of the writer than on the media as such; however, tweets (Twitter messages) tend to be more formal than chat messages in that they more often follow grammatical norms and use standard lexical items (Hu et al., 2013), while chats are more conversational (Paolillo, 1999), and hence less formal. Although social media often convey more ungrammatical text than more formal writings, Baldwin et al. (2013) show that the relative occurrence of non-standard syntax is fairly constant among many types of media, such as mails, tweets, forums, comments, and blogs, and argue that it should be tractable to develop NLP tools to process those, if focusing on English.

That is a large "if", though: first, the texts that we will discuss in this paper are not all in English, and — most importantly — not in one single language at all, but rather in a mix of languages, which clearly vastly complicates the issue of developing tools for these texts. Second, most previous research on social media text has focused on tweets, because of the ease of availability of Twitter; however, the conversational nature of chats tend to increase the level of code-mixing (Cárdenas-Claros and Isharyanti, 2009; Paolillo, 2011), so we will base our findings on data both from Twitter *and* from Facebook chats.

2.2 Code-Mixing and Tagging

There have been several efforts on social media text POS tagging in recent years, but almost exclusively on Twitter and mostly for English (Darling et al., 2012; Owoputi et al., 2013; Derczynski et al., 2013) and German (Rehbein, 2013; Neunerdt et al., 2014). Foster et al. (2011) introduce results for both POS tagging and parsing, but do not present a tool, and focus more on the parsing aspect. The two papers most similar to our work introduce the ARK tagger (Gimpel et al., 2011) and T-Pos (Ritter et al., 2011). The ARK tagger reaches 92.8% accuracy at token level, but uses a coarse, custom tagset. T-Pos is based on the Penn Treebank set and achieves an 88.4% token tagging accuracy. Neither paper reports sentence/whole tweet accuracy rates.

The first attempts at applying machine learning approaches to code-mixed language were by Solorio and Liu (2008a) who aimed to predict potential code alternation points, as a first step in the development of more accurate methods for processing code-mixed English-Spanish data. Only a few researchers have tried to tag code-mixed social media text: Solorio and Liu (2008b) addressed English-Spanish, while the English-Hindi mix was previously discussed by Vyas et al. (2014). Both used strategies based on combining the output of language-specific taggers, and we will utilize a similar solution in one of our experiments.

Turning to the specific problem of processing code-mixed Indian language data, Bhattacharja (2010) took a linguistic point of view on a particular type of complex predicates in Bengali that consist of an English word and a Bengali verb, in the light of different recent morphology models. Ahmed et al. (2011) noted that code-mixing and abbreviations add another dimension of transliteration errors of Hindi, Bengali and Telugu data when trying to understand the challenge of designing back-transliteration based input method editors. Mukund and Srihari (2012) proposed a tagging method that helps select words based on POS categories that strongly reflect Urdu-English code-mixing behavior. Das and Gambäck (2013) reported the first social media Indian code-mixing data (Bengali-Hindi-English), while Barman et al. (2014a) noted that character n-grams, part-of-speech, and lemmas were useful features for automatic language identification. Barman et al. (2014b) also carried out word-level classification experiments using a simple dictionary-based method. Bali et al. (2014) pointed out that structural and discourse linguistic analysis is required in order to fully analyse this type of code-mixing.

3 Data Collection and Annotation

For this work we have collected text both from Facebook and Twitter, initially 4,435 raw tweets and 1,236 Facebook posts. The tweets were on various 'hot' topics (i.e., topics that are currently being discussed in news, social media, etc.) and collected with the Java-based Twitter API,[1] while the Facebook posts were collected from campus-related university billboard postings (IIT Bombay Facebook Confession page).[2] The Facebook messages typically consist of a longer post (a "confession") followed by shorter, chat-like comments. The confessions are about "naughty" things that students have done on campus, and mainly concern cheating on exams or sex-related events.

Tokens	Facebook	Twitter	Total
Hindi	4.17	48.48	21.93
English	75.61	22.24	54.22
Universal	16.53	21.54	18.54
Named entity	2.19	6.70	3.99
Acronym	1.46	0.88	1.12
Mixed	0.02	0.08	0.05
Undefined	0.01	0.07	0.03

Table 1: Token Level Language Distribution (%) ('Universal' stands for punctuation marks, etc.)

3.1 Corpus

1,106 of the collected messages were randomly selected for manual annotation: 552 Facebook posts and 554 tweets. 20.8% of those messages are monolingual. Token level distribution of the corpus is reported in Table 1. Note that the Facebook messages are predominantly written in English, while the tweets mainly are in Hindi.

Utterance boundaries were manually inserted into the messages by two annotators, who initially agreed on 71% of the utterance breaks. After discussions and corrections, the agreement between the annotators was 94% and the resulting corpus has in total 2,583 utterances (1,181 from Twitter and 1,402 from Facebook), with 1,762 (68.2%) being monolingual. The sharp decrease in code-mixing when measured at the utterance level rather than message level shows the importance of the utterance boundary insertion, an issue we will get back to in Section 5.

Tokenization is an important preprocessing step which is difficult for social media text due to its

[1] http://twitter4j.org/

[2] www.facebook.com/Confessions.IITB

noisy nature. We used the CMU tokenizer,[3] which is a sub-module of the CMU Twitter POS tagger (Gimpel et al., 2011). Although the CMU tokenizer was originally developed for English, empirical testing showed that it works reasonably well also for the Indian languages.

3.2 Part-of-Speech Tagsets

We experimented with both coarse-grained and fine-grained tagsets, utilizing the fine-grained set during annotation. As can be seen in Table 2, this tagset includes both the Twitter specific tags introduced by Gimpel et al. (2011) and a set of POS tags for Indian languages that combines the IL-POST tags (Baskaran et al., 2008), the tags developed by the Central Institute of Indian Languages (LDCIL), and those suggested by the Indian Government's Department of Information Technology (TDIL),[4] that is, an approach similar to that taken for Gujarati by Dholakia and Yoonus (2014). The coarse-grained tagset instead combines Gimpel et al.'s Twitter specific tags with Google's Universal Tagset (Petrov et al., 2011).[5] The mapping between our fine-grained tagset and the Google Universal Tagset is also shown in Table 2.

3.3 Comparing Corpora Complexity

The error rates for various language processing applications would be expected to be higher for more complex code-mixed text. When comparing different code-mixed corpora to each other, it is thus desirable to have a measurement of the level of mixing between languages. Kilgarriff (2001) discusses various statistical measures that can be used to compare corpora more objectively, but all those measures presume the corpora to be monolingual.

In Das and Gambäck (2014) we instead suggested a *Code-Mixing Index*, CMI, to document the frequency of languages in a corpus, which we will use here as well. In short, the measure is defined as: if an utterance only contains language independent tokens, its CMI is zero; for other utterances, the CMI is calculated by counting the number of words belonging to the most frequent language in the utterance ($max\{w_i\}$) and dividing this by the total number of tokens (n) minus the number of language independent tokens (u):

Category	Type	Description
Noun (G_N)	N_NN	Common Noun
	N_NNV	Verbal Noun
	N_NST	Spatio-temporal
	N_NNP	Proper Noun
Pronoun (G_PRP)	PR_PRP	Personal
	PR_PRL	Relative
	PR_PRF	Reflexive
	PR_PRC	Reciprocal
	PR_PRQ	Wh-Word
Verb (G_V)	V_VM	Main
	V_VAUX	Auxiliary
Adjective (G_J)	JJ	Adjective
Adverb (G_R)	RB_ALC	Locative Adverb
	RB_AMN	Adverb of Manner
Demonstrative (G_PRP)	DM_DMD	Absolute
	DM_DMI	Indefinite
	DM_DMQ	Wh-word
	DM_DMR	Relative
Quantifier (G_SYM)	QT_QTF	General
	QT_QTC	Cardinal
	QT_QTO	Ordinal
Particles (G_PRT)	RP_RPD	Default
	RP_NEG	Negation
	RP_INTF	Intensifier
	RP_INJ	Interjection
Residual (G_X)	RD_RDF	Foreign Word
	RD_SYM	Symbol
	RD_PUNC	Punctuation
	RD_UNK	Unknown
	RD_ECH	Echo Word
Conjunction, Pre- & Postposition	CC	Conjunction
	PSP	Pre-/Postposition
Numeral	&	Numeral
Determiner	DT	Determiner
Twitter-Specific (Gimpel *et al.* 2011) (G_X)	@ ~	At-mention Re-Tweet/discourse
	E U #	Emoticon URL or email Hashtag

Table 2: POS Tagset

$$\text{CMI} = \begin{cases} 100 \times [1 - \frac{max\{w_i\}}{n-u}] & : n > u \\ 0 & : n = u \end{cases}$$

which means that for mono-lingual utterances, CMI $= 0$ (since then $max\{w_i\} = n - u$).

In Gambäck and Das (2014), we describe the index further and suggest that a factor that could be included in the index is the number of code alternation points (P) in an utterance, since a higher

[3] www.ark.cs.cmu.edu/TweetNLP/

[4] www.ldcil.org/Download/Tagset/LDCIL/6Hindi.pdf resp. www.tdil-dc.in/tdildcMain/articles/780732DraftPOSTagstandard.pdf

[5] The Google Universal Tagset defines the following twelve POS tags: G_N (nouns), G_V (verbs), G_J (adjectives), G_R (adverbs), G_PRP (pronouns), G_DT (determiners and articles), G_PRE (prepositions and post-positions), G_NUM (numerals), G_CONJ (conjunctions), G_PRT (particles), G_SYM (punctuation marks) and G_X (a catch-all for other categories such as abbreviations or foreign words).

CMI	Facebook	Twitter	P
Range	(%)	(%)	(avg.)
$[0]$	84.80	48.19	0.00
$(0, 10]$	4.49	3.11	1.75
$(10, 20]$	4.42	15.39	1.91
$(20, 30]$	3.49	14.38	2.37
$(30, 40]$	1.71	11.10	2.65
$(40, 100)$	1.06	7.14	2.70

Table 3: Code Mixing and Code Alternation

number of switches in an utterance arguably increases its complexity. However, that paper does not extend the CMI with code alternation points, and in the following we just separately report the average number of code alternation points. Details for our corpus are given in Table 3, based on CMI rages and code alternation point distributions.

Testing the idea that the Code-Mixing Index can describe the complexity of code-switched corpora, we used it to compare the level of language mixing in our English–Hindi corpus (in total, and each of the Facebook and Twitter parts in isolation) to that of the English-Hindi corpus of Vyas et al. (2014), the Dutch-Turkish corpus introduced by Nguyen and Doğruöz (2013), and the corpora used in the 2014 shared tasks at FIRE and EMNLP.[6] Table 4 shows the average CMI values for these corpora, both over all utterances and over only the utterances having a non-zero CMI (i.e., the utterances that contain some code-mixing). The last column of the table gives the fraction of mixed utterances in the respective corpora.

4 Part-of-Speech Tagging Experiments

This section discusses the actual tagging experiments, starting by describing the features used for training the taggers, and then reporting the results of using four different machine learning methods Finally, we contrast this with a strategy based on using a combination of language specific taggers.

4.1 Features

Feature selection plays a key role in supervised POS tagging. The important features for the POS tagging task have been identified based on the different possible combinations of available word and tag contexts. The features include the *focus word* (the current word), and its prefixes and suffixes from one-to-four letters (so four features each). Other features account for the previous word, the following word, whether the focus word starts with a digit or not, the previous word's POS tag, and the focus word's language tag.

Languages		CMI		P	Mixed
		avg.	mixed	(avg.)	(%)
EN-HI	FB+TW	13.38	21.86	2.33	61.21
	FB	3.67	13.24	2.50	27.71
	TW	23.06	24.38	2.28	94.58
	Vyas	2.54	14.82	2.15	20.68
DU-TR		21.48	26.46	4.43	26.55
FIRE	EN-GU	5.47	25.47	1.56	21.47
	EN-KN	14.29	21.43	5.50	66.66
	EN-ML	18.74	25.33	2.47	74.00
	EN-TA	25.00	37.50	3.00	66.66
	EN-BN	29.37	32.27	0.91	91.00
	EN-HI	19.32	24.41	4.89	79.14
EMNLP	EN-ES	6.93	24.13	0.31	28.70
	EN-ZH	10.15	19.43	0.97	52.75
	EN-NE	18.28	25.11	1.42	72.79
	AR-AR	4.41	25.60	0.17	17.21

Table 4: Code-Mixing in Various Corpora

Most of the features are self explanatory and quite obvious in POS tagging experiments, so we will only elaborate on prefix/suffix feature extraction: There are two different ways in which the focus word's suffix/prefix information can be used. The first and naïve one is to take a fixed length (say, n) suffix/prefix of the current and/or the surrounding word(s). If the length of the corresponding word is less than or equal to $n - 1$ then the feature value is not defined. The feature value is also not defined if the token itself is a punctuation symbol or contains any special symbol or digit.

The second and more helpful approach is to modify the feature to be binary or multiple valued. Variable length suffixes of a word can be matched with predefined lists of useful suffixes for different classes. Heuristic character extraction is generally not easy to motivate in theoretical linguistic terms, but the use of prefix/suffix information serves the practical purpose well for POS tagging of highly inflected languages, such as the Indian ones.

4.2 Machine Learning-based Taggers

We experimented with applying four machine learning-based classification algorithms to the

[6] The EMNLP corpora mix English with Spanish, Mandarin Chinese and Nepalese. The last EMNLP corpus is dialectal: Standard Arabic mixed with Egyptian (ARB-ARZ). The FIRE corpora mix English with Hindi, Gujarati, Bengali, Kannada, Malayalam and Tamil,. Note that the EN-KN and EN-TA corpora are very small (55 resp. 29 words), while EN-TA and EN-ML are only partially and inconsistently annotated, so those are not reliable as basis for comparison.

CMI	CRF		NB		SMO		RF	
Range	FG	CG	FG	CG	FG	CG	FG	CG
[0]	73.2	79.4	33.9	36.8	37.9	45.6	73.9	79.0
(0, 10]	64.0	71.5	36.0	40.1	39.0	45.9	68.7	75.3
(10, 20]	61.5	70.0	35.2	31.8	35.6	38.2	61.5	68.4
(20, 30]	60.4	68.0	33.3	42.0	36.3	46.6	58.2	67.3
(30, 40]	62.6	69.8	37.7	43.4	37.9	49.2	60.0	66.5
(40, 100)	64.5	71.1	39.2	44.3	39.0	49.3	62.4	67.6
avg.	64.3	71.6	35.8	39.7	37.6	45.8	64.1	70.6

Table 5: F_1 scores by CMI range distribution

Features	FG (F_1)	CG (F_1)
current word	62.0	67.7
+ next word	60.3	65.2
+ previous word	56.8	62.1
+ prefix	69.4	76.0
+suffix	72.2	78.9
+ start_with_digit	72.1	79.1
+ current_word_lang	73.3	79.8
+ prev_word_pos	73.3	79.8

Table 6: Feature Ablation for the RF-based Tagger

task: Conditional Random Fields (CRF), Sequential Minimal Optimization (SMO), Naïve Bayes (NB), and Random Forests (RF). For the CRF we used the MIRALIUM[7] implementation, while the other three were the implementations in WEKA.[8]

Table 5 reports performance after 5-fold cross validation of all the ML methods on the complete dataset (2,583 utterances), using both fine-grained (FG) and coarse-grained (CG) tagsets. As can be seen, Random Forests and CRF invariably gave the highest F scores (weighted average over all tags) on both tagsets, while SMO and Naïve Bayes consistently performed much worse. The difference between RF and CRF is not significant at the 99%-level in a paired two-tailed Student t-test.

To better understand the code-mixed POS tagging problem, we investigated which features are most important by performing feature ablation for RF-based tagger on the part of the corpus with CMI > 0. The feature ablation is reported in Table 6, with performance given by weighted average F-measure. As we see, including the previous or following word actually makes the performance decrease, while the other features contribute roughly the same to increase performance.

We then tested system performance on various number of code alternation points. Error rates at the alternation points are reported in Table 7, with the first column showing from which language the code alteration is taking place. The results indicate that all the ML methods have more problems with HI-EN alternation. A plausible reason is that most of the corpus is English mixed in Hindi, so the induced systems are biased towards Hindi syntactic patterns. More experiments are needed to better recognize which language is mixing into which, and to make the systems account for this; currently we are working on language modelling of code-mixed text for this purpose.

From	CRF		NB		SMO		RF	
To	FG	CG	FG	CG	FG	CG	FG	CG
EN-HI	12.4	9.0	21.2	18.9	21.2	17.8	12.1	8.5
HI-EN	5.4	5.6	19.2	18.1	18.2	16.6	4.8	4.6

Table 7: Error Rates (%) by Alternation Direction

4.3 Combining Language Specific Taggers

Solorio and Liu (2008b) proposed a simple but elegant solution of tagging code-mixed English-Spanish text twice — once each with a tagger for each language — and then combining the output of the language specific taggers to find the optimal word-level labels.

The reported accuracy of the combined tagger of Solorio and Liu (2008b) was 89.72%, when word-level languages were known. They used the Penn Treebank tagset, which is comparable to our fine-grained tagset, but since the CMI value for their English-Spanish corpus is not known, it is hard to compare the performance figures.

However, Vyas et al. (2014) followed the same strategy as Solorio and Liu (2008b), reporting an accuracy of 74.87%, also given that the word-level languages were known. They used the Google Universal Tagset and therefore in this way is comparable to our coarse-grained tagset, although (as can be seen in Table 4) the English-Hindi corpus used by Vyas et al. (2014) is far less mixed (has an average CMI of 2.54) than our English-Hindi corpus (with an average CMI of 13.38), plausibly justifying a higher POS tagging accuracy.

Word sequence plays a major role for syntactic formation as well as semantic meaning of the language, and could as such strongly influence POS tagging. The combination tagging strategy could potentially break the word sequences, so using language specific taggers is not necessarily the optimal approach; still, we have also carried out

[7] code.google.com/p/miralium/

[8] www.cs.waikato.ac.nz/ml/weka/

CMI Range	FG (%)	CG (%)
$[0]$	77.4	83.5
$(0, 10]$	69.5	75.9
$(10, 20]$	56.2	64.3
$(20, 30]$	59.9	68.2
$(30, 40]$	60.0	67.1
$(40, 100)$	66.4	72.8
avg.	64.9	72.0

Table 8: Accuracy of the Combination Tagger

experiments based on a similar language specific tagger combination, both for reasons of comparison and since the combination strategy is appealing in its straight-forward applicability.

The word-level language identifier of Barman et al. (2014b) (with a reported accuracy of 95.76%) was used to mark up our English-Hindi bilingual corpus with language tags for Hindi and English. To tag the Hindi tokens we then used the SNLTR[9] POS tagger, while CMU's ARK tagger was used to tag English and language independent tokens (i.e., universals, named entities, and acronyms).

As can be seen in Table 8, this gave an average accuracy of 71.97% on the coarse-grained tagset, marginally lower than the tagger's performance reported by Vyas et al. (2014), but compatible with the performance of the Random Forests and Conditional Random Field taggers described above. On the fine-grained tagset the tagger combination gave an average accuracy of 64.91%, also compatible with using the individual taggers.

5 Discussion

The ML-based taggers failed to out-perform the language specific combination tagger. One reason for this can be that the corpora used for training the machine learners is too small. Another reason might be that the Unknown Word Ratio (UWR) in these types of social media is very high. Unknown words typically cause problems for POS tagging systems (Giménez and Màrquez, 2004; Nakagawa et al., 2001). Our hypothesis was that the unknown word ratio increases with CMI. To test this, we calculated UWR on our English-Hindi corpus using both 10 folds and 5 folds, as shown in Table 9, getting numbers around 20% overall, with about 17% for the Facebook subpart and 29% for the Twitter part, supporting the hypothesis that the unknown word ratio indeed is high in these types of texts.

Folds	Facebook	Twitter	Total
5	17.03	29.95	20.49
10	16.68	29.27	19.79

Table 9: Average Unknown Word Ratios

Working with social media text has several other fundamental challenges. One of these is sentence and paragraph boundary detection (Reynar and Ratnaparkhi, 1997; Sporleder and Lapata, 2006), which definitely is a problem in its own right — and obviously extra difficult in the social media context. The importance of obtaining the correct utterance splitting is shown by the level of code-mixing dropping in our corpus when measuring it at utterance level rather than message level. For example, the following tweet could be considered to consist of two utterances U1 and U2:

(1) listening to Ishq Wala Love (From " Student of the Year ") The DJ Suketu Lounge Mix

U1 listening to Ishq Wala Love (From " Student of the Year ")

U2 The DJ Suketu Lounge Mix

But one can also argue that this is one utterance only: even though the "The" is capitalized, it just starts a subordinate clause. In more formal language, it probably would have been written as:

(2) Listening to Ishq Wala Love (from "Student of the Year"), the DJ Suketu Lounge Mix.

Utterance boundary detection for social media text is thus a challenging problem in itself, which was not discussed in detail by Gimpel et al. (2011) or Owoputi et al. (2013). The main reason might be that those works were on tweets, that are limited to 140 characters, so even if the whole tweet is treated as one utterance, POS tagging results will not be strongly affected. However, when working with Facebook messages, we found several long posts, with a high number of code alternation points (6–8 alternation points are very common).

Automatic utterance boundary detection for social media text clearly demands separate solution mechanisms. In this work we have manually marked the utterance boundaries, but see Read et al. (2012) and López and Pardo (2015) for suggestions for how to address the problem.

[9] http://nltr.org/snltr-software/

6 Conclusion and Future Work

The paper has aimed to put the spotlight on the issues that make code-mixed text challenging for language processing. We report work on collecting, annotating, and measuring the complexity of code-mixed English-Hindi social media text (Twitter and Facebook posts), as well as experiments on automatic part-of-speech tagging of these corpora, using both a coarse-grained and a fine-grained tagset. Four machine learning algorithms were applied to the task (Conditional Random Fields, Sequential Minimal Optimization, Naïve Bayes, and Random Forests), and compared to a language specific combination tagger. The RF-based tagger performed best, but only marginally better than the combination tagger and the one based on CRFs.

There are several possible avenues that could be further explored on NLP for code-mixed texts, for example, transliteration, utterance boundary detection, language identification, and parsing. We are currently working on language modelling of code-mixed text to recognize which language is mixing into which. Language modelling has not before been applied to code-mixed POS tagging, but code-switched language models have previously been integrated into speech recognisers, although mostly by naïvely interpolating between monolingual models. Li and Funng (2014) instead obtained a code-switched language model by combining the matrix language model with a translation model from the matrix language to the mixed language. In the future, we also wish to explore language modelling on code-mixed text in order to address the problems caused by unknown words.

Acknowledgements

Thanks to all the researchers who have made their datasets and tools available: the organisers of the shared tasks on code-switching at EMNLP 2014 and in transliteration at FIRE 2014, Dong Nguyen (University of Twente The Netherlands), Seza Doğruöz (Tilburg University, The Netherlands), Monojit Choudhury and Kalika Bali (both at Microsoft Research India), and Utsab Barman (Dublin City University, Ireland).

References

Umair Z Ahmed, Kalika Bali, Monojit Choudhury, and Sowmya VB. 2011. Challenges in designing input method editors for Indian languages: The role of word-origin and context. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand, November. AFNLP. Workshop on Advances in Text Input Method.

Jannis Androutsopoulos. 2011. Language change and digital media: a review of conceptions and evidence. In Tore Kristiansen and Nikolas Coupland, editors, *Standard Languages and Language Standards in a Changing Europe*, pages 145–159. Novus, Oslo, Norway,

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan, October. AFNLP.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing *ya* mixing?": An analysis of English-Hindi code mixing in Facebook. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 116–126, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014a. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 13–23, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching.

Utsab Barman, Joachim Wagner, Grzegorz Chrupała, and Jennifer Foster. 2014b. DCU-UVT: Word-level language classification with code-mixed data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 127–132, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching.

Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Monojit Choudhury, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha, and KVS Subbarao. 2008. A common parts-of-speech tagset framework for Indian languages. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1331–1337, Marrakech, Marocco, May. ELRA.

Shishir Bhattacharja. 2010. Bengali verbs: a case of code-mixing in Bengali. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 75–84, Sendai, Japan, November.

Mónica Stella Cárdenas-Claros and Neny Isharyanti. 2009. Code switching and code mixing in internet chatting: between 'yes', 'ya', and 'si' a case study. *Journal of Computer-Mediated Communication*, 5(3):67–78.

William M. Darling, Michael J. Paul, and Fei Song. 2012. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic Bayesian HMM. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–9, Avignon, France, April. ACL. Workshop on Semantic Analysis in Social Media.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text: The last language identification frontier? *Traitement Automatique des Langues*, 54(3):41–64.

Amitava Das and Björn Gambäck. 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*, pages 169–178, Goa, India, December.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing*, pages 198–206, Hissar, Bulgaria, September.

Purva S. Dholakia and M. Mohamed Yoonus. 2014. Rule based approach for the transition of tagsets to build the POS annotated corpus. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(7):7417–7422, July.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June. ACL.

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef Van Genabith. 2011. #hardtoparse: POS tagging and parsing the twitterverse. In *Proceedings of the 25th National Conference on Artifical Intelligence*, pages 20–25, San Fransisco, California, August. AAAI. Workshop On Analyzing Microtext.

Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 1–7, Goa, India, December. 1st Workshop on Language Technologies for Indian Social Media.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 168–176, Lisbon, Portugal, May. ELRA.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2: short papers, pages 42–47, Portland, Oregon, June. ACL.

Yuhen Hu, Kartik Talamadupula, and Subbarao Kambhampati. 2013. *Dude, srsly?*: The surprisingly formal nature of Twitter's language. In *Proceedings of the 7th International Conference on Weblogs and Social Media*, Boston, Massachusetts, July. AAAI.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th International Conference on Computational Linguistics*, pages 145–150, Prague, Czechoslovakia, July. ACL.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Ying Li and Pascale Funng. 2014. Code switch language modeling with functional head constraint. In *Proceedings of the 2014 International Conference on Acoustics, Speech and Signal Processing*, pages 4946–4950, Florence, Italy, May. IEEE.

Roque López and Thiago A.S. Pardo. 2015. Experiments on sentence boundary detection in user-generated web content. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Proceedings of the 16th International Conference*, pages 357–368, Cairo, Egypt, March. Springer.

Smruthi Mukund and Rohini K. Srihari. 2012. Analyzing Urdu social media for sentiments using transfer learning with controlled translations. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–8, Atlanta, Georgia, June. ACL. 2nd Workshop on Language in Social Media.

Tetsuji Nakagawa, Taku Kudoh, and Yuji Matsumoto. 2001. Unknown word guessing and part-of-speech tagging using support vector machines. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, pages 325–331, Tokyo, Japan.

Melanie Neunerdt, Michael Reyer, and Rudolf Mathar. 2014. Efficient training data enrichment and unknown token handling for POS tagging of non-standardized texts. In Josef Ruppenhofer and Gertrud Faaß, editors, *Proceedings of the 12th Edition of the KONVENS Conference*, pages 186–192, Hildesheim, Germany, October. Universitätsverlag Hildesheim.

Dong Nguyen and A. Seza Doğruöz. 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, Seattle, Washington, October. ACL.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. ACL.

John Paolillo. 1996. Language choice on soc.culture.punjab. *Electronic Journal of Communication*, 6(3), June.

John Paolillo. 1999. The virtual speech community: Social network and language variation on IRC. *Journal of Computer-Mediated Communication*, 4(4), June.

John Paolillo. 2011. "conversational" codeswitching on usenet and internet relay chat. *Language@Internet*, 8(article 3), June.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2011. A universal part-of-speech tagset. *CoRR*, abs/1104.2086.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 985–994, Mumbai, India, December. ACL. Poster.

Ines Rehbein. 2013. Fine-grained POS tagging of German tweets. In *Proceedings of the 25th International Conference on Language Processing and Knowledge in the Web*, pages 162–175, Darmstadt, Germany, September. Springer.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 803–806, Washington, DC, April. ACL.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, August. ACL.

Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii, October. ACL.

Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii, October. ACL.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching.

Caroline Sporleder and Mirella Lapata. 2006. Broad coverage paragraph segmentation across languages and domains. *ACM Transactions on Speech and Language Processing*, 3(2):1–35, July.

Drahomíra Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 763–771, Athens, Greece, March. ACL.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 974–979, Doha, Qatar, October. ACL.