POS Tagging of English-Hindi Code-Mixed Social Media Content

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

Spandana Gella∗
Xerox Research Centre Europe
spandanagella@gmail.com

Jatin Sharma Kalika Bali Monojit Choudhury
Microsoft Research India
{jatin.sharma,kalikab,monojitc}@microsoft.com

Abstract

Code-mixing is frequently observed in user generated content on social media, especially from multilingual users. The linguistic complexity of such content is compounded by presence of spelling variations, transliteration and non-adherence to formal grammar. We describe our initial efforts to create a multi-level annotated corpus of Hindi-English code-mixed text collated from Facebook forums, and explore language identification, back-transliteration, normalization and POS tagging of this data. Our results show that language identification and transliteration for Hindi are two major challenges that impact POS tagging accuracy.

1 Introduction

Code-Switching and Code-Mixing are typical and well-studied phenomena of multilingual societies (Gumperz, 1964; Auer, 1984; Myers-Scotton, 1993; Danet and Herring, 2007; Cardenas-Claros and Isharyanti, 2009). Linguists differentiate between the two, where Code-Switching is juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems (Gumperz, 1982), and Code-Mixing (CM) refers to the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language (Myers-Scotton, 1993). The first example in Fig. 1 features CM where English words are embedded in a Hindi sentence, whereas the second example shows codeswitching. Here, we will use CM to imply both. Work on computational models of CM have been few and far between (Solorio and Liu, 2008a; Solorio and Liu, 2008b; Nguyen and Dogruoz, 2013), primarily due to the paucity of CM data in conventional text-corpora which makes data-intensive methods hard to apply. Solorio and Liu (2008a) in their work on English-Spanish CM use models built on smaller datasets to predict valid switching points to synthetically generate data from monolingual corpora, and in another work (2008b) describe parts-of-speech (POS) tagging of CM text.

CM though typically observed in spoken language is now increasingly more common in text, thanks to the proliferation of the Computer Mediated Communication channels, especially social media like Twitter and Facebook (Crystal, 2001; Herring, 2003; Danet and Herring, 2007; Cardenas-Claros and Isharyanti, 2009). Social media content is tremendously important for studying trends, reviews, events, human-behaviour as well as linguistic analysis, and therefore in recent times has spurred a lot of interest in automatic processing of such data. Nevertheless, CM on social media has not been studied from a computational aspect. Moreover, social media content presents additional challenges due to contractions, non-standard spellings and non-grammatical constructions. Furthermore, for languages written in scripts other than Roman, like Hindi, Bangla, Japanese, Chinese and Arabic, Roman transliterations are typically used for representing the words (Sowmya et al., 2010). This can prove a challenge for language identification and segregation of the two languages.

In this paper, we describe our initial efforts to POS tag social media content from English-Hindi (henceforth En-Hi) bilinguals while trying to address the challenges of CM, transliteration and non-standard spelling, as well as lack of annotated data. POS tagging is one of the fundamental pre-processing steps for NLP, and while there

∗ This work was done during authors’ internship at Microsoft Research India.
have been works on POS tagging of social media
data (Gimpel et al., 2011; Owoputi et al., 2013)
and of CM (Solorio and Liu, 2008b), but we do
not know of any work on POS tagging of CM
text from social media that involves transliteration.
The salient contributions of this work are in for-
malizing the problem and related challenges for
processing of En-Hi social media data, creation
of an annotated dataset and some initial exper-
iments for language identification, transliteration,
normalization and POS tagging of this data.

2 Corpus Creation

For this study, we collected data from Face-
book public pages of three celebrities: Amitabh
Bachchan, Shahrukh Khan, Narendra Modi, and
the BBC Hindi news page. All these pages are
very popular with 1.8 to 15.5 million “likes”. A to-
tal of 40 posts were manually selected from these
pages, which were published between 22nd – 28th
October 2013. The posts having a long thread of
comments (50+) were preferred, because CM and
non-standard usage of language is more common
in the comments. We shall use the term post to re-
fer to either a post or a comment. The corpus thus
created has 6,983 posts and 113,578 words. The
data was semi-automatically cleaned and format-
ted. The user names were removed for anonymity,
but the names appearing in comments, which are
mostly of celebrities, were retained.

2.1 Annotation

There are various interesting linguistic as well as
socio-pragmatic features (e.g., user demograph-
ics, presence of sarcasm or humor, polarity) for
which this corpus could be annotated because CM
is influenced by both linguistic as well as extra-
linguistic features. However, initial attempts at
such detailed and layered annotation soon revealed
the resource-intensiveness of the task. We, thus,
scaled down the annotation to the following four
layers:

Matrix: The posts are split into contiguous
fragments of words such that each fragment has a
unique matrix language (either En or Hi). The
matrix language is defined as the language which
governs the grammatical relation between the con-
stituents of the utterance. Any other language
words that are nested into the matrix constitute the
eMBEDDED language(s). Usually, matrix language
can be assigned to clauses or sentences.

Word origin: Every word is marked for its ori-
gin or source language, En or Hi, depending on
whether it is an English or Hindi word. Words that
are of neither Hindi nor English origin are marked
as Ot or Other. Here, we assume that code-mixing
does not happen at sublexical levels, as it is un-
common in this data; Hi and En have a sim-
pler inflectional morphology and thus, sub-lexical
mixing though present (e.g., computer has an
En root - computer and a Hi plural marker
on) is relatively less common. In languages with
richer morphology and agglutination, like Bangla
and most Dravidian languages, more frequent sub-
lexical mixing may be observed. Also note that
words are borrowed extensively between Hi and
En such that certain English words (e.g., bus,
party, vote etc) are no longer perceived as English
words by the Hindi speakers. However, here we
will not distinguish between CM and borrowing,
and such borrowed English words have also been
labeled as En words.

Normalization/Transliteration: Whenever the
word is in a transliterated form, which is often the
case for the Hi words, it is labeled with the in-
tended word in the native script (e.g., Devanagari
for Hi). If the word is in native script, but uses
a non-standard spelling, it is labeled with the cor-
correct standard spelling. We call this the spelling
normalization layer.

Parts-of-Speech (POS): Finally, each word is
also labeled with its POS. We use the Universal
POS tagset proposed by Petrov et al. (2011) which
has 12 POS tags that are applicable to both En
and Hi. The POS labels are decided based on the
function of a word in the context, rather than a
decontextualized lexical category. This is an im-
portant notion, especially for CM text, because of-
ten the original lexical category of an embedded
word is lost in the context of the matrix language,
and it plays the role of a different lexical category.
Though the Universal POS tagset does not pre-
scribe a separate tag for Named Entities, we felt
the necessity of marking three different kinds of
NEs - people, location and organization, because
almost every comment has one or more NEs and
strictly speaking word origin does not make sense
for these words.

Annotation Scheme: Fig. 1 illustrates the an-
notation scheme through two examples. Each
post is enclosed within <s></s> tags. The
matrices within a post are separated by the
<matrix></matrix> tags which take the matrix
language as an argument. Each word is anno-
tated for POS, and the language (/E or /H for En or Hi respectively) only if it is different from the language of the matrix. In case of non-standard spelling in English, the correct spelling is appended as “soul”, while for the Hindi words, the correct Devanagari transliteration is appended. The NEs are marked with the tags P (person), L (location) or O (organization) and multiword NEs are enclosed within square brackets “[[]]”.

A random subsample of 1062 posts consisting of 10171 words were annotated by a linguist who is a native speaker of Hi and proficient in En. The annotations were reviewed and corrected by two experts linguists. During this phase, it was also observed that a large number of comments were very short, typically an eulogism of their favorite celebrity and hence were not interesting from a linguistic point of view. For our experiments, we removed all posts that had fewer than 5 words. The resulting corpus had 381 comments/posts and 4135 words.

2.2 CM Distribution

Most of the posts (93.17%) are in Roman script, and only 2.93% were in Devanagari. Around 3.5% of the posts contain words in both the scripts (typically a post in Devanagari with hashtags or urls in Roman script), and a very small fraction of the text (0.4% of comments/posts and 0.6% words) was in some other script. The fraction of words present in Roman and Devanagri scripts are 80.76% and 15.32% respectively, which shows that the Devanagari posts are relatively longer than the Roman posts. Due to their relative rarity, the posts containing words in Devanagari or any other script were not considered for annotation.

In the annotated data, 1102 sentences are in a single matrix (398 Hi, 698 En and 6 Ot) and in 45 posts there is at least one switch of matrix (mostly between Hi and En). Thus, 4.2% of the data shows code-switching. This is a strict definition of code-switching; if we consider a change in matrix within a conversation thread as a code-switch, then in this data all the threads exhibit code-switching. However, out of the 398 comments in Hi-matrix, 23.37% feature CM (i.e., they have at least one or more non-Hi (or rather, almost always En) words embedded. On the other hand, only 7.34% En-matrix comments feature CM (again almost always with Hi). Thus, a total of 17.2% comments/posts, which contains a quarter of all the words in the annotated corpus, feature either CM or code-switching or both. We also note that more than 40% words in the corpus are in Hi or other Indian languages, but written in Roman script; hence, they are in transliterated form. See (Bali et al., 2014) for an in-depth discussion on the characteristics of the CM data.

This analysis demonstrates the necessity of CM and transliterated text processing in the context of Indian user-generated social media content. Perhaps, the numbers are not too different for such content generated by the users of any other bilingual and multilingual societies.

3 Models and Experiments

POS tagging of En-Hi code-mixed data requires language identification at both word and matrix level as well back-transliteration of the text into
Table 1: Confusion matrix, precision and recall of the language identification module.

<table>
<thead>
<tr>
<th>Actual Label</th>
<th>Predicted Label</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi</td>
<td>1057</td>
<td>515</td>
</tr>
<tr>
<td>En</td>
<td>45</td>
<td>2023</td>
</tr>
<tr>
<td>Precision</td>
<td>0.959</td>
<td>0.797</td>
</tr>
</tbody>
</table>

Table 1 shows the token (word) level confusion matrix for the language identification task on our dataset. The language labels of 84.6% of the tokens were correctly predicted by the system. As can be seen from the Table, the precision for predicting Hi is high, whereas that for En is low. This is mainly due to the presence of a large number of contracted and distorted Hi words in the dataset, e.g., h for hai (Fig. 1), which were tagged as En by our system because the training examples had no contracted Hi words, but short and non-conventional spellings were in plenty in the En training examples as those were extracted from the SMS corpus.

3.2 Normalization

In our dataset, if a word is identified as Hi, then it must be back-transliterated to Devanagari script so that any off-the-shelf Hindi POS tagger can be used. We used the system by Gella et al. (Gella et al., 2013) for this task, which is part rule-based and part statistical. The system was trained on the 35000 unique transliteration pairs extracted from Hindi song lyrics (Gupta et al., 2012). This corpus has a reasonably wide coverage of Hindi words, and past researchers have also shown that transliteration does not require a very large amount of training data. Normalization of the En text was not needed because the POS tagger (Owoputi et al., 2013) could handle unnormalized text.

3.3 POS tagging

Solorio and Liu (2008b) describes a few approaches to POS-tagging of code-switched Spanish text, all of which primarily relies on two monolingual taggers and certain heuristics to combine the output from the two. One of the simpler heuristics is based on language identification, where the POS tag of a word is the output of the monolingual tagger of the language in which the word is. In this initial study, we apply this basic idea for POS tagging of CM data. We divide the text (which is already sentence-separated) into contiguous maximal chunks of words which are in the same language. Then we apply a Hi POS tagger to the Hi chunks, and an En POS tagger to the En chunks.
<table>
<thead>
<tr>
<th>Model</th>
<th>LI</th>
<th>HN</th>
<th>Tagger</th>
<th>Hi Acc.</th>
<th>En Acc.</th>
<th>Total Acc.</th>
<th>Hi CA</th>
<th>En CA</th>
<th>Total CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>K</td>
<td>K</td>
<td>Standard</td>
<td>75.14</td>
<td>81.91</td>
<td>79.02</td>
<td>27.34</td>
<td>39.67</td>
<td>34.05</td>
</tr>
<tr>
<td>1b</td>
<td>K</td>
<td>K</td>
<td>Twitter</td>
<td>75.14</td>
<td>82.66</td>
<td>79.62</td>
<td>27.34</td>
<td>35.74</td>
<td>31.91</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>NK</td>
<td>Twitter</td>
<td>65.61</td>
<td>81.73</td>
<td>74.87</td>
<td>17.58</td>
<td>33.77</td>
<td>26.38</td>
</tr>
<tr>
<td>3</td>
<td>NK</td>
<td>NK</td>
<td>Twitter</td>
<td>44.74</td>
<td>80.68</td>
<td>65.39</td>
<td>40.00</td>
<td>13.17</td>
<td>25.00</td>
</tr>
</tbody>
</table>

Table 2: POS Tagging accuracies for the different models. K=Known, NK = Not Known. LI = Language labels, HN = Hindi normalized forms, Acc. = Token level accuracy, CA = Chunk level accuracy.

We use a CRF++ based POS tagger for Hi, which is freely available from [http://nltr.org/snltr-software/](http://nltr.org/snltr-software/). For En, we use the Twitter POS tagger (Owoputi et al., 2013). It also has an inbuilt tokenizer and can work directly on unnormalized text. This tagger has been chosen because Facebook posts and comments are more Twitter-like. We also use the Stanford POS Tagger (Toutanova et al., 2003) which, unlike the Twitter POS Tagger, has not been tuned for Twitter-like text. These taggers use different tagsets - the ILPOST for Hi (Sankaran et al., 2008) and Penn-TreeBank for En (Marcus et al., 1993). The output tags are appropriately mapped to the smaller Universal tagset (Petrov et al., 2011).

### 3.4 Experiments and Results

We conducted three different experiments as follows. In the first experiment, we assume that we know the language identities and normalized/transliterated forms of the words, and only do the POS tagging. This experiment gives us an idea of the accuracy of POS tagging task, if normalization, transliteration and language identification could be done perfectly. We conduct this experiments with two different En POS taggers: the Stanford POS tagger which is trained on formal English text (Model 1a) and the Twitter POS tagger (Model 1b). In the next experiment (Model 2), we assume that only the language identity of the words are known, but for Hindi we apply our model to generate the back transliterations. For English, we apply the Twitter POS tagger directly because it can handle unnormalized social media text. The third experiment (Model 3) assumes that nothing is known. So language identifier is first applied, and based on the language detected, we apply the Hi transliteration module, and Hi POS tagger, or the En tagger. This is the most challenging and realistic setting. Note that the matrix information is not used in any of our experiments, though it could be potentially useful for POS tagging and could be explored in future.

Table 2 gives a summary of the four models along with the POS tagging accuracies (in %). It shows token level as well as chunk level accuracies (CA), i.e., what percentage of chunks have been correctly POS tagged. As can be seen, Hi POS tagging has relatively low accuracies than En POS tagging at word level for all cases. This is primarily due to the errors of the transliteration module, which in turn, is because the transliteration does not address spelling contractions. This is also reflected in the drop in the accuracies for the case where LI is unknown. The very low CA for En for model 3 is primarily because some of the Hi chunks are incorrectly identified as En by the language identification module (see Table 1). However, the gradual drop of token and chunk level accuracies from model 1 to model 3 clearly shows the effect of gradual error accumulation from each of the modules. We observe that Nouns were usually confused most with Verbs and vice versa, while the Adj were mostly confused with Nouns, Pronouns with Determiners, and Adpositions with Conjunctions.

### 4 Conclusion

This is a work in progress. We have identified normalization and transliteration as two very challenging problems for En-Hi CM text. Joint modelling of language identification, normalization, transliteration as well as POS tagging is expected to yield better results. We plan to continue our work in that direction, specifically for conversational text in social media in a multilingual context. CM is a common phenomenon found in all bilingual and multilingual societies. The issue of transliteration exist for most of the South Asian languages as well as many other languages such as Arabic and Greek, which use a non-Roman based script (Gupta et al., 2014). The challenges and issues identified in this study are likely to hold for many other languages as well, which makes this a very important and globally prevalent problem.
References


Olutobi Owoputi, Brendan OConnor, 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 NAACL.


