All that is English may be Hindi: Enhancing language identification through automatic ranking of likeliness of word borrowing in social media

[1]**Jasabanta Patro**, [2]**Bidisha Samanta**, [2]**Saurabh Singh**,
[2]**Abhipsa Basu**, [2]**Prithwish Mukherjee**, [3]**Monojit Choudhury**, [4]**Animesh Mukherjee**
[1,2,4] Indian Institute of Technology Kharagpur, India – 721302
[6]Microsoft Research India, Bangalore – 560001
[1]jasabantapatro@iitkgp.ac.in, [3]monojitc@microsoft.com, [4]animeshm@cse.iitkgp.ernet.in

Abstract

In this paper, we present a set of computational methods to identify the likeliness of a word being borrowed, based on the signals from social media. In terms of Spearman's correlation values, our methods perform more than two times better (~ 0.62) in predicting the borrowing likeliness compared to the best performing baseline (~ 0.26) reported in literature. Based on this likeliness estimate we asked annotators to re-annotate the language tags of foreign words in predominantly native contexts. In 88% of cases the annotators felt that the foreign language tag should be replaced by native language tag, thus indicating a huge scope for improvement of automatic language identification systems.

1 Introduction

In social media communication, multilingual people often switch between languages, a phenomenon known as *code-switching* or *code-mixing* (Auer, 1984). This makes *language identification and tagging*, which is perhaps a prerequisite for almost all other language processing tasks that follow, a challenging problem (Barman et al., 2014). In code-mixing people are subconsciously aware of the foreign origin of the code-mixed word or the phrase. A related but linguistically and cognitively distinct phenomenon is *lexical borrowing* (or simply, *borrowing*), where a word or phrase from a foreign language say L_2 is used as a part of the vocabulary of native language say L_1. For instance, in Dutch, the English word "sale" is now used more frequently

than the Dutch equivalent "uitverkoop". Some English words like "shop" are even inflected in Dutch as "shoppen" and heavily used. While it is difficult in general to ascertain whether a foreign word or phrase used in an utterance is borrowed or just an instance of code-mixing (Bali et al., 2014), one tell tale sign is that only proficient multilinguals can code-mix, while even monolingual speakers can use borrowed words because, by definition, these are part of the vocabulary of a language. In other words, just because an English speaker understands and uses the word "tortilla" does not imply that she can speak or understand Spanish. A borrowed word from L_2 initially appears frequently in speech, then gradually in print media like newspaper and finally it loses its origin's identity and is used in L_1 resulting in an inclusion in the dictionary of L_1 (Myers-Scotton, 2002; Thomason, 2003). Borrowed words often take several years before they formally become part of L_1 dictionary. This motivates our research question "is early-stage automatic identification of likely to be borrowed words possible?". This is known to be a hard problem because (i) it is a socio-linguistic phenomenon closely related to acceptability and frequency, (ii) borrowing is a dynamic process; new borrowed words enter the lexicon of a language as old words, both native and borrowed, might slowly fade away from usage, and (iii) it is a population level phenomenon that necessitates data from a large portion of the population unlike standard natural language corpora that typically comes from a very small set of authors. Automatic identification of borrowed words in social media content (SMC) can improve language tagging by recommending the tagger to tag the language of the borrowed words as L_1 instead

of L_2. The above reasons motivate us to resort to the social media (in particular, Twitter), where a large population of bilingual/multilingual speakers are known to often tweet in code-mixed colloquial languages (Carter et al., 2013; Solorio et al., 2014; Vyas et al., 2014; Jurgens et al., 2017; Rijhwani et al., 2017). We designed our methodology to work for any pair of languages L_1 and L_2 subject to the availability of sufficient SMC. In the current study, we consider Hindi as L_1 and English as L_2.

The main stages of our research are as follows:
Metrics to quantify the likeliness of borrowing from social media signals: We define three novel and closely similar metrics that serve as social signals indicating the likeliness of borrowing. We compare the likeliness of borrowing as predicted by our model and a baseline model with that from the ground truth obtained from human judges.
Ground truth generation: We launch an extensive survey among 58 human judges of various age groups and various educational backgrounds to collect responses indicating if each of the candidate foreign word is likely borrowed.
Application: We randomly selected some words that have a high, low and medium borrowing likeliness as predicted by our metrics. Further, we randomly selected one tweet for each of the chosen words. The chosen words in almost all of these tweets have L_2 as their language tag while a majority of the surrounding words have a tag L_1. We asked expert annotators to re-evaluate the language tags of the chosen words and indicate if they would prefer to switch this tag from L_2 from L_1.

Finally, our key results are outlined below:
1. We obtained the Spearman's rank correlation between the ground-truth ranking and the ranking based on our metrics as ~ 0.62 for all the three variants which is more than double the value (~ 0.26) if we use the most competitive baseline (Bali et al., 2014) available in the literature.
2. Interestingly, the responses of the judges in the age group below 30 seem to correspond even better with our metrics. Since language change is brought about mostly by the younger population, this might possibly mean that our metrics are able to capture the early signals of borrowing.
3. Those users that mix languages the least in their tweets present the best signals of borrowing in case they do mix the languages (correlation of our metrics estimated from the tweets of these users

with that of the ground truth is ~ 0.65).
4. Finally, we obtain an excellent re-annotation accuracy of 88% for the words falling in the surely borrowed category as predicted by our metrics.

2 Related work

In linguistics *code-mixing* and *borrowing* have been studied under the broader scope of language change and evolution. Linguists have for a long time focused on the sociological and the conversational necessity of borrowing and mixing in multilingual communities (see Auer (1984) and Muysken (1996) for a review). In particular, Sankoff et al. (1990) describes the complexity of choosing features that are indicative of borrowing. This work further showed that it is not always true that only highly frequent words are borrowed; nonce words could also be borrowed along with the frequent words. More recently, (Nzai et al., 2014) analyzed the formal conversation of Spanish-English multilingual people and found that code mixing/borrowing is not only restricted to daily speech but is also prevalent in formal conversations. (Hadei, 2016) showed that phonological integration could be evaluated to understand the phenomenon of word borrowing. Along similar lines, (Sebonde, 2014) showed morphological and syntactic features could be good indicators for numerical borrowings. (Senaratne, 2013) reported that in many languages English words are likely to be borrowed in both formal and semi-formal text.

Mixing in computer mediated communication and social media: (Sotillo, 2012) investigated various types of code-mixing in a corpora of 880 SMS text messages. The author observed that most often mixing takes place at the beginning of a sentence as well as through simple insertions. Similar observations about chat and email messages have been reported in (Bock, 2013; Negrón, 2009). However, studies of code-mixing with Chinese-English bilinguals from Hong Kong (Li, 2009) and Macao (San, 2009) brings forth results that contrast the aforementioned findings and indicate that in these societies code-mixing is driven more by linguistic than social motivations.

Recently, the advent of social media has immensely propelled the research on code-mixing and borrowing as a dynamic social phenomena. (Hidayat, 2012) noted that in Facebook, users mostly preferred inter-sentential mixing and showed that 45% of the mixing originated from

real lexical needs, 40% was used for conversations on a particular topic and the rest 5% for content clarification. In contrast, (Das and Gambäck, 2014) showed that in case of Facebook messages, intra-sentential mixing accounted for more than half of the cases while inter-sentential mixing accounted only for about one-third of the cases. In fact, in the First Workshop on *Computational Approaches to Code Switching* a shared task on code-mixing in tweets was launched and four different code-mixed corpora were collected from Twitter as a part of the shared task (Solorio et al., 2014). Language identification task has also been handled for English-Hindi and English-Bengali code-mixed tweets in (Das and Gambäck, 2013). Part-of-speech tagging have been recently done for code-mixed English-Hindi tweets (Solorio and Liu, 2008; Vyas et al., 2014).

Diachronic studies: As an aside, it is interesting to note that the availability of huge volumes of timestamped data (tweet streams, digitized books) is now making it possible to study various linguistic phenomena quantitatively over different timescales. For instance, (Sagi et al., 2009) uses latent semantic analysis for detection and tracking of changes in word meaning, whereas (Frermann and Lapata, 2016) presents a Bayesian approach for the same problem. (Peirsman et al., 2010) presents a distributed model for automatic identification of lexical variation between language varieties. (Bamman and Crane, 2011) discusses a method for automatically identifying word sense variation in a dated collection of historical books . (Mitra et al., 2014) presents a computational method for automatic identification of change in word senses across different timescales. (Cook et al., 2014) presents a method for novel sense identification of words over time.

Despite these diverse and rich research agendas in the field of code-switching and lexical dynamics, there has not been much attempt to quantify the likeliness of borrowing of foreign words in a language. The only work that makes an attempt in this direction is (Bali et al., 2014), which is described in detail in Sec 3.1. One of the primary challenges faced by any quantitative research on lexical borrowing is that borrowing is a social phenomenon, and therefore, it is difficult to identify suitable indicators of such a lexical diffusion process unless one has access to a large population-level data. In this work, we show for the first time

how certain simple and closely related signals encoding the language usage of social media users can help us construe appropriate metrics to quantify the likeliness of borrowing of a foreign word.

3 Methodology

In this section, we present the baseline metric and propose three new metrics that quantify the likeliness of borrowing.

3.1 Baseline metric

Baseline metric – We consider the $log(\frac{F_{L_2}}{F_{L_1}})$ value proposed in (Bali et al., 2014) as the baseline metric. Here F_{L_2} denotes the frequency of the L_1 transliterated form of the word w in the standard L_1 newspaper corpus. F_{L_1}, on the other hand, denotes the frequency of the L_1 translation of the word w in the same newspaper corpus. For our experiments discussed in the later sections, both the transliteration and the translation of the words have been done by a set of volunteers who are native L_1 speakers. The authors in (Bali et al., 2014) claim that the more positive the value of this metric is for a word w, the higher is the likeliness of its being borrowed. The more negative the value is, the higher are the chances that the word w is an instance of code-mixing.

Ranking – Based on the values obtained from the above metric for a set of target words, we rank these words; words with high positive values feature at the top of the rank list and words with high negative values feature at the bottom of the list. For two words having the same $log(\frac{F_{L_2}}{F_{L_1}})$ value, we resolve the conflict by assigning each of these the average of their two rank positions. In a subsequent section, we shall compare this rank list with the one obtained from the ground truth responses.

3.2 Proposed metric

In this section, we present three novel and closely related metrics based on the language usage patterns of the users of social media (in specific, Twitter). In order to define our metrics, we need all the words to be language tagged. The different tags that a word can have are: L_1, L_2, *NE* (Named Entity) and *Others*. Based on the word level tag, we also create a tweet level tag as follows:

1. L_1: Almost every word ($> 90\%$) in the tweet is tagged as L_1.
2. L_2: Almost every word ($> 90\%$) in the tweet is tagged as L_2.

3. CML_1: Code-mixed tweet but majority (i.e., > 50%) of the words are tagged as L_1.
4. CML_2: Code-mixed tweet but majority (i.e., > 50%) of the words are tagged as L_2.
5. *CMEQ*: Code-mixed tweet having very similar number of words tagged as L_1 and L_2 respectively.
6. *Code Switched*: There is a trail of L_1 words followed by a trail of L_2 words or vice versa.

Using the above classification, we define the following metrics:

Unique User Ratio (UUR) – The Unique User Ratio for word usage across languages is defined as follows:

$$UUR(w) = \frac{U_{L_1} + U_{CML_1}}{U_{L_2}} \qquad (1)$$

where U_{L_1} (U_{L_2}, U_{CML_1}) is the number of unique users who have used the word w in a L_1 (L_2, CML_1) tweet at least once.

Unique Tweet Ratio (UTR) – The Unique Tweet Ratio for word usage across languages is defined as follows:

$$UTR(w) = \frac{T_{L_1} + T_{CML_1}}{T_{L_2}} \qquad (2)$$

where T_{L_1} (T_{L_2}, T_{CML_1}) is the total number of L_1 (L_2, CML_1) tweets which contain the word w.

Unique Phrase Ratio (UPR) – The Unique Phrase Ratio for word usage across languages is defined as follows:

$$UPR(w) = \frac{P_{L_1}}{P_{L_2}} \qquad (3)$$

where P_{L_1} (P_{L_2}) is the number of L_1 (L_2) phrases which contain the word w. Note that unlike the definitions of UUR and UTR that exploit the word level language tags, the definition of UPR exploits the phrase level language tags.

Ranking – We prepare a separate rank list of the target words based on each of the three proposed metrics – UUR, UTR and UPR. The higher the value of each of this metric the higher is the likeliness of the word w to be borrowed and higher up it is in the rank list. In a subsequent section, we shall compare these rank lists with the one prepared from the ground truth responses.

4 Experiments

In this section we discuss the dataset for our experiments, the evaluation criteria and the ground truth preparation scheme.

4.1 Datasets and preprocessing

In this study, we consider code-mixed tweets gathered from Hindi-English bilingual Twitter users in order to study the effectiveness of our proposed metrics. The native language L_1 is Hindi and the foreign language L_2 is English. To bootstrap the data collection process, we used the language tagged tweets presented in (Rudra et al., 2016). In addition to this, we also crawled tweets (between Nov 2015 and Jan 2016) related to 28 hashtags representing different Indian contexts covering important topics such as sports, religion, movies, politics etc. This process results in a set of 811981 tweets. We language-tag (see details later in this section) each tweet so crawled and find that there are 3577 users who use mixed language for tweeting. Next, we systematically crawled the time lines of these 3577 users between Feb 2016 and March 2016 to gather more mixed language tweets. Using this two step process we collected a total of 1550714 distinct tweets. From this data, we filtered out tweets that are not written in romanized script, tweets having only URLs and tweets having empty content. Post filtering we obtained 725173, tweets which we use for the rest of the analysis. The datasets can be downloaded from http://cnerg.org/borrow

Language tagging: We tag each word in a tweet with the language of its origin using the method outlined in (Gella et al., 2013). *Hi* represents a predominantly Hindi tweet, *En* represents a predominantly English tweet, *CMH* (*CME*) represents code-mixed tweets with more Hindi (English) words, *CMEQ* represents code-mixed tweets with almost equal number of Hindi and English words and *CS* represents code-switched tweets (the number and % of tweets in each of the above six categories are presented in the supplementary material). Like the word level, the tagger also provides a phrase level language tag. Once again, the different tags that an entire phrase can have are: *En*, *Hi* and *Oth* (Other). The metrics defined in the previous section are computed using these language tags.

Newspaper dataset for the baseline: As we had discussed in the previous section, for the construction of the baseline ranking we need to resort to counting the frequency of the foreign words (i.e., English words) and their Hindi translations in a newspaper corpus as has been outlined in (Bali et al., 2014). For this purpose, we use the FIRE

dataset built from the Hindi Jagaran newspaper corpus[1] which is written in Devanagari script.

4.2 Target word selection

We first compute the most frequent foreign (i.e., English words) in our tweet corpus. Since we are interested in the frequency of the English word only when it appears as a foreign word we do not consider the (i) *Hi* tweets since they do not have any foreign word, (ii) *En* tweets since here the English words are not foreign words and the (iii) code-switched tweets. Based on the frequency of usage of English as a foreign word, we select the top 1000 English words. Removal of stop words and text normalization leaves beyond 230 nouns (see supplementary material for the list of words).

Final selection of target words: In language processing, context plays an important role in understanding different properties of a word. For our study, we also attempt to use the language tags as features of the context words for a given *target word*. Our hypothesis here is that there should exist classes of words that have similar context features and the likelihood of being borrowed in each class should be different. For example, when an English word is surrounded by mostly Hindi words it seems to be more likely borrowed. We present two examples in the box below to illustrate this.

Example I:
@****** Welcome. ***Film** jaroor dekhna.*
Nahi to injection ready hai.
Example II:
@*****_Trust @*****
Kuch to ache se karo sirji....
Har jagah bhaagte rehna is not a good ***thing***.

In *Example I* the English word "film" is surrounded by mostly Hindi words. On the other hand, in *Example II* the English word "thing" is surrounded mostly by English words. Note that the word "film" is very commonly used by Hindi monolingual speakers and is therefore highly likely to have been borrowed unlike the English word "thing" which is arguably an instance of mixing. This socio-linguistic difference seems to be very appropriately captured by the language tag of the surrounding words of these two words in the respective tweets. Based on this hypothesis

---

we arrange the 230 words into contextually similar groups (see supplementary material for the grouping details). Finally, using the baseline metric $log(\frac{F_E}{F_H})$ (E: English, H: Hindi), we proportionately choose words from these groups as follows:

Words with very high or very low values of $log(\frac{F_E}{F_H})$ *(hlws)* – we select words having the highest and the lowest values of $log(\frac{F_E}{F_H})$ from each of the context groups. This constitutes a set of 30 words. Note that these words are baseline-biased and therefore the metric should be able to discriminate them well.

Words with medium values of $log(\frac{F_E}{F_H})$ *(mws)* – we selected 27 words having not so high and not so low $log(\frac{F_E}{F_H})$ at uniformly at random.

Full set of words *(full)* – Thus, in total we selected 57 target words for the purpose of our evaluation. We present these words in the box below.

Baseline-biased words – 'thing', 'way', 'woman', 'press', 'wrong', 'well', 'matter', 'reason', 'question', 'guy', 'moment', 'week', 'luck', 'president', 'body', 'job', 'car', 'god', 'gift', 'status', 'university', 'lyrics', 'road', 'politics', 'parliament', 'review', 'scene', 'seat', 'film', 'degree'
Randomly selected words – 'people', 'play', 'house', 'service', 'rest', 'boy', 'month', 'money', 'cool', 'development', 'group', 'friend', 'day', 'performance', 'school', 'blue', 'room', 'interview', 'share', 'request', 'traffic', 'college', 'star', 'class', 'superstar', 'petrol', 'uncle'

4.3 Evaluation criteria

We present a four step approach for evaluation as follows. We measure (i) how well the UUR, UTR and UPR based ranking of the $hlws$ set, the mws set and the $full$ set correlate with the ground truth ranking (discussed in the next section) in comparison to the rank given by the baseline metric, (ii) how well the different rank ranges obtained from our metric align with the ground truth as compared to the baseline metric, (iii) whether there are some systematic effects of the age group of the survey participants on the rank correspondence, (iv) how our metrics if computed from the tweets of users who (a) rarely mix languages, (b) almost always mix languages and (c) are in between (a) and (b), align with the ground truth.

Rank correlation: We measure the standard Spearman's rank correlation (ρ) (Zar, 1972) pairwise between rank lists generated by (i) UUR (ii) UTR (iii) UPR (iv) baseline metric and the ground truth.

We shall describe the next four measurements

taking UUR as the running example. The same can be extended verbatim for the other two similar metrics.

Rank ranges: We split each of the three rank lists (UUR, ground truth and baseline) into five different equal-sized ranges as follows – (i) surely borrowed (SB) containing top 20% words from each list, (ii) likely borrowed (LB) containing the next 20% words from each list, (iii) borderline (BL) constituting the subsequent 20% words from each list, (iv) likely mixed (LM) comprising the next 20% words from each list and (v) surely mixed (SM) having the last 20% words from each rank list. Therefore, we have three sets of five buckets, one set each for UUR, the ground truth and the baseline based rank list.

Next we calculate the bucket-wise correspondence between (i) the UUR and the ground truth set and (ii) the baseline and the ground truth set in terms of standard *precision* and *recall* measures. For our purpose, we adapt these measures as follows.

G: ground truth bucket set, B_b: baseline bucket set, U_b: UUR bucket set;

$BS \in \{B_b, U_b\}$, T (type of bucket) = {SB, LB, BL, LM, SM};

b_t = words in type t bucket from BS, g_t = words in type t bucket from G, $t \in T$;

tp_t (no. of true positives) = $|b_t \cap g_t|$, fp_t (no. of false positives) = $|b_t - g_t|$, tn_t (no. of true negatives) = $|g_t - b_t|$;

Bucket-wise *precision* and *recall* therefore:

$precision(b_t) = \frac{tp_t}{fp_t + tp_t}$; $recall(b_t) = \frac{tp_t}{tn_t + tp_t}$

For a given set, we obtained the overall *macro precision* (*recall*) by averaging the *precision* (*recall*) values over the five buckets. For a given set, we also obtained the overall *micro precision* by first adding the true positives across all the buckets and then normalizing by the sum of the true and the false positives over all the buckets. We take an equivalent approach for obtaining the *micro recall*.

Age group effect: Here we construct two ground truth rank lists one using the responses of the participants with age below 30 (young population) and the other using the responses of the rest of the participants (elderly population). Next we repeat the above two evaluations considering each of the new ground truth rank lists.

Extent of language mixing: Here we divide all the 3577 users into three categories – (i) High (users who have more than 20% of tweets as code-mixed), (ii) Mid (users who have 7–20% of their tweets as code-mixed, and (iii) Low (users who have less than 7% of their tweets as code-mixed). We create three UUR based rank lists for each of these three user categories and respectively compare them with the ground truth rank list.

4.4 Ground truth preparation

Since it is very difficult to obtain a suitable ground truth to validate the effectiveness of our proposed ranking scheme, we launched an online survey to collect human judgment for each of the 57 *target words*.

Online survey We conducted the online survey[2] among 58 volunteers majority of whom were either native language(Hindi) speakers or had very high proficiency in reading and writing in that language. The participants were selected from different age groups and different educational backgrounds. Every participant was asked to respond to a multiple choice question about each of the 57 *target words*. Therefore, for every single *target word*, 58 responses were gathered. The multiple choice question had the following three options and the participants were asked to select the one they preferred the most and found more natural – (i) a Hindi sentence with the *target word* as the only English word, (ii) the same Hindi sentence in (i) but with the *target word* replaced by its Hindi translation and (iii) none of the above two options. There were no time restrictions imposed while gathering the responses, i.e., the volunteers theoretically had unlimited time to decide their responses for each *target word*.

Language preference factor For each *target word*, we compute a *language preference factor* (LPF) defined as ($Count_{En} - Count_{Hi}$), where $Count_{Hi}$ refers to the number of survey participants who preferred the sentence containing the Hindi translation of the *target word* while $Count_{En}$ refers to the number of survey participants who preferred the sentence containing the *target word* itself. More positive values of LPF denotes higher usage of *target word* as compared to its Hindi translation and therefore higher likeliness of the word being borrowed.

Ground truth rank list generation We generate the ground truth rank list based on the LPF score of a *target word*. The word with the highest value

---

[2]Survey portal: https://goo.gl/forms/L0kJm8BNMhRj0jA53

2259

Rank-List$_1$	Rank-List$_2$	$\rho - hlws$	$\rho - mws$	$\rho - full$
UUR	Ground truth	**0.67**	**0.64**	**0.62**
UTR	Ground truth	**0.66**	**0.63**	**0.63**
UPR	Ground truth	**0.66**	**0.64**	**0.62**
Baseline	Ground truth	0.49	0.14	0.26

Table 1: Spearman's rank correlation coefficient (ρ) among the different rank lists. Best result is marked in bold.

Bucket type	Ground truth bucket	Baseline bucket	UUR bucket
SB	11	11	11
LB	11	11	11
BL	12	12	12
LM	11	11	11
SM	12	12	12

Table 2: Number of words falling in each bucket of three bucket sets.

Bucket type	*prec./rec.* Baseline	*prec./rec.* UUR
SB	**0.27**	**0.27**
LB	0.09	**0.18**
BL	0.08	**0.33**
LM	0.18	**0.36**
SM	0.33	**0.50**

Table 3: Bucket-wise *precision/recall*. Best results are marked in bold.

of LPF appears at the top of the ground truth rank list and so on in that order. Tie breaking between *target words* having equal LPF values is done by assigning average rank to each of these words.

Age group based rank list: As discussed in the previous section, we prepare the age group based rank lists by first splitting the responses of the survey participants in two groups based on their age – (i) young population (age < 30) and (ii) elderly population (age ≥ 30). For each group we then construct a separate LPF based ranking of the *target words*.

5 Results

5.1 Correlation among rank lists

The Spearman's rank correlation coefficient (ρ) of the rank lists for the $hlws$ set, the mws set and the $full$ set according to the baseline metric, UUR, UTR and UPR with respect to the ground truth metric LPF are noted in table 1. We observe that for the $full$ set, the ρ between the rank lists obtained from all the three metrics UUR, UTR, and UPR with respect to the ground truth is ~ 0.62 which is more than double the ρ (~ 0.26) between the baseline and the ground truth rank list. This clearly shows that the proposed metrics are able to identify the likeliness of borrowing quite accurately and far better than the baseline. Further, a remarkable observation is that our metrics outperform the baseline metric even for the $hlws$ set that is baseline-biased. Likewise, for the mws set, our metrics outperform the baseline indicating a superior recall on arbitrary words. The ranking of the $full$ set of words obtained from the ground truth, the baseline and the UUR metric is available in the supplementary material.

We present the subsequent results for the $full$ set and the UUR metric. The results obtained for the other two metrics UTR and UPR are very similar and therefore not shown.

5.2 Rank list alignment across rank ranges

The number of *target words* falling in each bucket across the three rank lists are the same and are noted in table 2. Thus, the *precision* and *recall* as per the definition are also the same. The bucket-wise *precision/recall* for the baseline and UUR with respect to the ground truth are noted in table 3. We observe that while in the SB bucket both the baseline and UUR perform equally well, for all the other buckets UUR massively outperforms the baseline. This implies that for the case where the likeliness of borrowing is the strongest, the baseline does as good as UUR. However, as one moves down the rank list, UUR turns out to be a considerably better predictor than the baseline. The overall *macro* and *micro precision/recall* as shown in table 4 further strengthens our observation that UUR is a better metric than the baseline.

5.3 Age group based analysis

As already discussed earlier, we split the ground truth responses based on the age group of the survey participants. We split the responses into two groups – (i) young population (age < 30) and (ii) elderly population (age ≥ 30) so that there are almost equal number of responses in both the groups (see supplementary material for the exact distribution).

Rank correlation: The Spearman's rank correlation of UUR and the baseline rank lists with these

Measure	Baseline	UUR
Macro prec./rec.	0.19	**0.33**
Micro prec./rec.	0.19	**0.33**

Table 4: Overall *macro* and *micro precision/recall*. Best results are marked in bold.

Rank-List$_1$	Rank-List$_2$	ρ
Baseline	Ground-truth-Young	0.26
UUR	Ground-truth-Young	**0.62**
Baseline	Ground-truth-Elder	0.27
UUR	Ground-truth-Elder	**0.53**

Table 5: Spearman's rank correlation across the two age groups. Best results are marked in bold.

Bucket type	Young-Baseline p/r	Young-UUR p/r	Elder-Baseline p	Elder-UUR p	Elder-baseline r	Elder-UUR r
SB	**0.27**	**0.27**	0.27	**0.36**	0.25	**0.33**
LB	0.09	**0.18**	0.09	**0.18**	0.08	**0.17**
BL	0.08	**0.33**	**0.16**	0.08	**0.28**	0.14
LM	0.18	**0.36**	0.18	**0.45**	0.14	**0.35**
SM	0.33	**0.5**	**0.41**	0.25	**0.41**	0.25

Table 6: Bucket-wise *precision* (*p*)/*recall* (*r*) for UUR and the *baseline* metrics for the two new ground truths. Best results are marked in bold.

two ground truth rank lists are shown in table 5. Interestingly, the correlation between UUR rank list and the young population ground truth is better than the elderly population ground truth. This possibly indicates that UUR is able to predict recent borrowings more accurately. However, note that the UUR rank list has a much higher correlation with both the ground truth rank lists as compared to the baseline rank list.

Rank ranges: Table 6 shows the bucket-wise *precision* and *recall* for UUR and the baseline metrics with respect to two new ground truths. For the young population once again the number of words in each bucket for all the three sets is the same thus making the values of the *precision* and the *recall* same. In fact, the *precision*/*recall* for this ground truth is exactly same as in the case of the original ground truth.

In contrast, when we consider the ground truth based on the responses of the elderly population, the number of words across the different buckets are different across the three sets. In this case, we observe that the *precision*/*recall* values are better for the UUR metric in SB, LB and LM buckets.

Finally, the overall *macro* and *micro precision* and *recall* for both the age groups are noted in table 7. Once again, for both the young and the elderly population based ground truths, the *macro* and *micro precision* and *recall* values for the UUR metric are higher compared to that of the baseline.

5.4 Extent of language mixing

As mentioned earlier, we divide the set of 3577 users into three categories. The Spearman's cor-

Measure	Young-Baseline	Young-UUR	Elder-Baseline	Elder-UUR
Mac. pre.	0.19	**0.33**	0.22	**0.27**
Mac. rec.	0.19	**0.33**	0.23	**0.25**
Mic. pre./rec.	0.19	**0.33**	0.23	**0.26**

Table 7: Overall *macro* and *micro precision* and *recall* for the two new ground truths. Best results are marked in bold.

Bucket	Number of users	ρ
High	302	0.52
Mid	744	0.60
Low	2531	**0.65**

Table 8: Spearman's correlation between UUR and the ground truth in the different user buckets. Best results are marked in bold.

relation between UUR and the ground truth for each of these buckets are given in table 8. As we can see, for Low bucket the ρ value is maximum. This points to the fact that the signals of borrowing is strongest from the users who rarely mix languages.

6 Re-annotation results

In order to conduct the re-annotation experiments we performed the following. To begin with, we ranked all the 230 English nouns in non-increasing order of their UUR values. We then randomly selected 20 words each having (i) high UUR (top 20%) values (call TOP), (ii) low UUR (bottom 20%) values (call BOT), and (iii) middle UUR (middle 20%) values (call MID). This makes a total of 60 words. Using this word list we extracted one tweet each that contained the (foreign) word from this list along with all other words in the tweet tagged in Hindi (H_{all}). We similarly prepared another such list of 60 words and extracted one tweet each in which most of the other words were tagged in Hindi (H_{most}).

We presented the selected words and the corresponding tweets to a set of annotators and asked them to annotate these selected words once again. Over all the words, we calculated the mean ($\mu_{E \to H}$) and the standard deviation ($\sigma_{E \to H}$) of the fraction of cases where the annotators altered the tag of the selected word from English to Hindi. The average inter-annotator agreement (Fleiss, 1971) for our experiments was found to be as high as 0.64. For the words in the TOP list, the fraction of times the tag is altered is **0.91** (**0.85**) with an inter-annotator agreement of 0.84 (0.80) for the

Word-rank	Context	$\mu_{E \to H}$	$\sigma_{E \to H}$
TOP	H_{all}	**0.91**	0.15
TOP	H_{most}	**0.85**	0.23
MID	H_{all}	0.58	0.28
MID	H_{most}	0.61	0.34
BOT	H_{all}	0.13	0.18
BOT	H_{most}	0.16	0.21

Table 9: Re-annotation results

H_{all} (H_{most}) category. In other words, on average, in as high as 88% of cases the annotators altered the tags of the words that are highly likely to be borrowed (i.e., TOP) in a largely Hindi context (i.e., H_{all} or H_{most}). Table 9 shows the fractional changes for all the other possible cases. An interesting observation is that the annotators rarely flipped the tags for the words in the BOT list (i.e., the sure mixing cases) in either of the H_{all} or the H_{most} contexts. These results strongly support the inclusion of our metric in the design of future automatic language tagging tasks.

7 Discussion and conclusion

In this paper, we proposed a few new metrics for estimating the likliness of borrowing that rely on signals from large scale social media data. Our best metric is two-fold better than the previous metric in terms of the accuracy of the prediction. There are some interesting linguistic aspects of borrowing as well as certain assumptions regarding the social media users that have important repercussions on this work and its potential extensions, which are discussed in this section.

Types of borrowing: Linguists define broadly three forms of borrowing, (i) *cultural*, (ii) *core*, and (iii) *therapeutic* borrowings. In *cultural borrowing*, a foreign word gets borrowed into native language to fill a lexical gap. This is because there is no equivalent native language word present to represent the same foreign word concept. For instance, the English word 'computer' has been borrowed in many Indian languages since it does not have a corresponding term in those languages[3]. In *core borrowing*, on the other hand, a foreign word replaces its native language translation in the native language vocabulary. This occurs due to overwhelming use of the foreign word over native language translation as a matter of prestige, ease of use etc. For example, the English word 'school'

has become much more prevalent than its corresponding Hindi translation '*vidyalaya*' among the native Hindi speakers. Finally, therapeutic borrowing refers to borrowing of words to avoid taboo and homonomy in the native language. In this paper, although we did not perform any category based studies, most of our focus was on core borrowing.

Language of social media users: We assumed that if a user is predominantly using Hindi words in a tweet then the chances of him/her being a native Hindi speaker should be high, since, while the number of English native speakers in India is 0.02%, the number of Hindi native speakers is 41.03%[4]. This assumption has also been made in earlier studies (Rudra et al., 2016). Note that even if a user is not a native Hindi speaker but a proficient (or semi-proficient) Hindi speaker, the main results of our analysis should hold. For instance, consider two foreign words 'a' and 'b'. If 'a' is frequently borrowed in the native language, then the proficient speaker would also tend to borrow 'a' similar to a native speaker. Even if due to lack of adequate native vocabulary, the non-native speaker borrows the word 'b' in some cases, these spurious signals should get eliminated since we are making an aggregate level statistics over a large population.

Future directions: It would be interesting to understand and develop theoretical justification for the metrics. Further, it would be useful to study and classify various other linguistic phenomena closely related to core borrowing, such as: (i) *loanword*, where a form of a foreign word and its meaning or one component of its meaning gets borrowed, (ii) *calques*, where a foreign word or idiom is translated into existing words of native language, and (iii) *semantic loan*, where the word in the native language already exists but an additional meaning is borrowed from another language and added to existing meaning of the word.

Finally, we would also like to incorporate our findings into other standard tasks of multilingual IR and multilingual speech synthesis (for example to render the appropriate native accent to the borrowed word).

---

[3]Words for "computer" were coined in many Indian languages through morphological derivation from the term for "compute", however, none of these words are used in either formal or informal contexts.

[4]http://bit.ly/2ufOvea

References

P. Auer. 1984. *Bilingual Conversation*. John Benjamins.

K. Bali, J. Sharma, M. Choudhury, and Y. Vyas. 2014. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *First workshop on Computational approaches to code-switching, EMNLP*, page 116.

David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 1–10. ACM.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. *EMNLP 2014*, 13.

Z. Bock. 2013. Cyber socialising: Emerging genres and registers of intimacy among young south african students. *Language Matters: Studies in the Languages of Africa*, 44(2):68–91.

S. Carter, W. Weerkamp, and M. Tsagkias. 2013. Micro-blog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *COLING*, pages 1624–1635.

A. Das and B. Gambäck. 2013. code-mixing in social media text: The last language identification frontier? *TAL*, 54(3):41–64.

A. Das and B. Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. In *ICON*, pages 169–178.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

S. Gella, J. Sharma, and K. Bali. 2013. Query word labeling and back transliteration for indian languages: Shared task system description. *FIRE Working Notes*, 3.

M. Hadei. 2016. Single word insertions as code-switching or established borrowing? *International Journal of Linguistics*, 8(1):14–25.

T. Hidayat. 2012. An analysis of code switching used by facebookers (a case study in a social network site). In *BA thesis*. English Education Study Program, College of Teaching and Education (STKIP), Bandung, Indonesia.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. *ACL 2017*.

D. C. S. Li. 2009. Cantonese-english code-switching research in hong-kong: a y2k review. *World Englishes*, 19(3):305–322.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. *arXiv preprint arXiv:1405.4392*.

P. Muysken. 1996. Code-switching and grammatical theory. In *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 177–198. Cambridge University Press.

C. Myers-Scotton. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press.

R. G. Negrón. 2009. Spanish-english code-switching in email communication. *Language@Internet*, 6(3).

V. E. Nzai, Y-L. Feng, M. R. Medina-Jiménez, and J. Ekiaka-Oblazamengo. 2014. Understanding code-switching & word borrowing from a pluralistic approach of multilingualism.

Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. *ACL 2017*.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111. Association for Computational Linguistics.

H. K. San. 2009. Chinese-english code-switching in blogs by macao young people. In *MSc. thesis*. University of Edinburgh.

D. Sankoff, S. Poplack, and S. Vanniarajan. 1990. The case of the nonce loan in tamil. *Language variation and change*, 2(01):71–101.

R. Y. Sebonde. 2014. Code-switching or lexical borrowing: Numerals in chasu language of rural tanzania. *Journal of Arts and Humanities*, 3(3):67.

C D. W. Senaratne. 2013. Borrowings or code mixes: The presence of lone english nouns in mixed discourse.

T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Gohneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang, and P. Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *EMNLP First Workshop on Computational Approaches to Code Switching*, pages 62–72.

T. Solorio and Y. Liu. 2008. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics.

S. Sotillo. 2012. Ehhhh utede hacen plane sin mi???:@ im feeling left out :(form, function and type of code switching in sms texting. In *ICAME*, pages 309–310.

S. G. Thomason. 2003. Contact as a source of language change. *The handbook of historical linguistics*, pages 687–712.

Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury. 2014. POS tagging of english-hindi code-mixed social media content. In *EMNLP*, volume 14, pages 974–979.

J. H. Zar. 1972. Significance testing of the spearman's rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580.