Feasibility of Leveraging Crowd Sourcing for the Creation of a Large Scale Annotated Resource for Hindi English Code Switched Data: A Pilot Annotation

Mona Diab
Center for Computational Learning Systems
Columbia University, New York
mdiab@ccls.columbia.edu

Ankit Kamboj
Computer Science Dept.
Columbia University, New York
ak3171@columbia.edu

Abstract

Linguistic code switching (LCS) occurs when speakers mix multiple languages in the same speech utterance. We find LCS pervasively in bilingual communities. LCS poses a serious challenge to Natural Language and Speech Processing. With the ubiquity of informal genres online, LCS is emerging as a very widespread phenomenon. This paper presents a first attempt at collecting and annotating a large repository of LCS data. We target Hindi English (Hinglish) LCS. We investigate the feasibility of leveraging crowd sourcing as a means for annotating the data on the word level. This paper briefly explains the setup of the experiment and data collection. It also presents statistics representing agreements among annotators over different possible categories of Hinglish words and analyzes the confidence with which a code switched word can be annotated in the correct category by humans.

1 Introduction

Linguistic Code switching (LCS) is the term used to describe a common practice among bilingual speakers of a given language pair in which the speakers switch back and forth between their common languages. This phenomenon is dominantly observed in inhabitants of countries like India where Hindi is a common first language (L1) and English acts as a second language (L2) among native Hindi speakers. For example, the following Hindi sentence with code switches to English is a seamless example of North Indian conversation: ***Uske*** *communication* ***ki wajah se hi*** *project successful* ***hua hai***. *(Project has become successful because of his excellent communication.)* LCS occurs both inter-sentential and intra-sentential.

LCS occurs in all genres of communication for such speakers, including spoken conversation, email, online chat rooms, blogs and newsgroups. Thus, it seriously impacts attempts to process these exchanges computationally, for the purposes of automatic translation, speech recognition, and information extraction, inter alia (Solorio and Liu, 2008a; Solorio and Liu, 2008b).

With increasing interest in LCS, there is need for large annotated LCS corpora which can support the needs of computational as well as theoretical research. This paper presents one experiment where a corpus of code switched sentences is annotated for identifying code switch points using crowd sourcing methods. The data collection serves as the first attempt at creating a repository for LCS data. Also the annotations of LCS points will shed light into the nature of this phenomenon and will be an initial building block for the development of interesting analytical and predictive models for automatic LCS processing systems. It is widely accepted that LCS actually follows a certain pattern and that it does not occur randomly. Several studies in sociolinguistics and theoretical linguistics have investigated this issue however on a small scale (Poplack, 2001; Myers-Scotton, 1993).

2 Hindi and Hinglish

Hindi is the national language of India and native language of many parts of the country. It has continuously been impacted by varied languages and dialects of the country, the most influential of which is English. English expanded its roots into India from the time Britain occupied the country. It was initially the language of the elite upper class but as the education system became widespread, English spread across the whole country. With the proliferation of scientific advancements in the English speaking world and India's race for technology acquisition, we note that English has almost

Proceedings of the 9th Workshop on Asian Language Resources, pages 36–40,
Chiang Mai, Thailand, November 12 and 13, 2011.

become an Indian language. In fact, Indians from different parts of India who speak mutually unintelligible native Indian languages use English as the bridge language to communicate. The pervasiveness of English coupled with the Hindi education throughout the country led to rapid development of **Hinglish**, a term coined to describe the use of Hindi and English words in the same utterance, Hinglish LCS.[1] Since Hindi is a morphologically rich language, we even often observe LCS occurring on the morphological level. Hinglish has become a widespread phenomenon (as a language in and of itself even) used by Indians in different parts of the world. It is obvious that the context switches from Hindi to English are very frequent and some words that are borrowed such as *Thank You*, *Please*, *Crazy* are almost Hindi words, as they have become part of the Indian native lexicon. One important reason that smooth switch can occur between Hindi and English is that words from any of these languages can fill the lexical gaps in the sentence of the other. It is important to point out that LCS is beyond nonce and borrowing, the phenomenon in Hinglish is that of significant amounts of words and chunks are switched back and forth in the same utterance, it is not a matter of isolated borrowed words that are highly frequent in the Hindi lexicon.

Our paper attempts to describe an initial large scale collection of LCS data and annotate it on the word/token level. Several linguistic studies have investigated Hinglish on a theoretical level (Bhatt, 1997; Joshi, 1985) as well as socio-pragmatic level as in the work of Bhatt and Bolonyai (2008). The studies suggest that LCS occurs in a systematic manner. However to our knowledge no large collection of LCS data for Hinglish exists, let alone detailed annotations for such a collection. Our initial attempt is to fill this gap such that it would be of utility to both the theoretical linguistics as well as the computational processing fields.

3 Corpus Collection

We needed content where the matrix language was Hindi with frequent code switches to English. Modern Hindi novels are rich sources for such content as they use Hinglish frequently. The content of two sites: *www.hindinovels.net* and *www.abhivyakti-hindi.org* were crawled using perl scripts and broken into sentences to develop the corpus. Some Hindi sentences with no CS were mixed with these sentences to prepare an optimal blend of sentences for annotations. The final corpus consisted of 10500 sentences comprising 193285 tokens.

4 Experiment Setup

Amazon Mechanical Turk (AMT) is a marketplace to host surveys where **requesters** host some questions which are answered by **workers**, aka turkers. It has been widely accepted that the use of crowd sourcing techniques for the collection of data annotations is a worthwhile effort (Snow et al., 2008). The benefit of using crowd sourcing lies in a rapid collection cycle, sometimes at the expense of quality. Hence the challenge lies in designing and simplifying the task and presenting it to lay people in generic terms. But also setting performance metrics for accepting such annotations. We carried out our experiments on AMT where we asked the turkers to identify each word in a sentence as one of the following categories:[2]

1. Hindi- *aaya*(came),*gaya*(went),*hum*(we)
2. English- usual English words, for example, *eat, grin, happy*
3. Foreign Proper Name- *John*,*Stella*, *IBM*
4. Indian proper Name- *Ramesh*,*Ganesh*, *Anjali*
5. Unknown- Any word which can not be classified into any of the above categories

The experiment was set up as a survey with three Hinglish sentences on one page. Each of such pages is termed a Human Intelligence Task (HIT) and a collection of HITs is termed a task on AMT. Our collection of 10500 sentences was divided into 7 tasks, each task containing 500 HITs with 3 sentences each. Each word in a sentence had a drop down list containing the above options associated with it, with the default option being Hindi. The AMT turkers then marked each word in the sentence as one of the options above. A minimum of two turkers were allowed to work on a single HIT or the same set of 3 sentences in order to allow overlap for agreements/disagreements on same set of words. Accordingly all the data was at

[1] For the purposes of this paper, we are not interested in the inter sentential LCS.

[2] For this pilot annotation, we did not include the more complex annotation of mixed Hinglish morphology. We decided to postpone that annotation for a later phase.

least doubly annotated.

A subset of HITs (10% of the corpus size) was gold annotated by a native bilingual speaker of Hindi and English. We designed the set up of the HITs such that for any given turker at least one sentence in a HIT overlapped with a gold annotation. Then the turker whose sampled HIT annotations agreed with the gold annotation less than 95% were discarded. Initially, 136 turkers submitted the results, out of which 85 turkers scored above the set 95% threshold. The HITs that were rejected were resubmitted to AMT for re-annotation. With resubmission results, 8 more turkers were added as they scored above the 95% threshold bringing the total number of turkers to 93. Accordingly, the overall data was annotated by 93 turkers, 10% of the overall 10500 sentences is three way annotated with gold annotation and by two turkers.

5 Experiment Results and Statistics

In this section we present detailed results on the collected annotations. We calculate inter-turker agreements based on how many times a turker agreed on a category within the same HIT with the other turkers who co-annotated the same HIT. The results for turkers were then aggregated to find the total number of agreements for each category. The resulting confusion matrix is shown in Table 1. The legend for the table is as follows:

h- Hindi
e- English
f- Foreign Proper Name
i- Indian Proper Name
u- Unknown

Each cell of the confusion matrix corresponds to agreement counts for any turker aggregately with respective co-turkers.

The following detailed statistics show the percentage classification agreement among the co-turkers in different categories for the majority annotated class on the word level. As mentioned above, each HIT was annotated by two turkers. In our detailed statistics, we observe the number of times two turkers agreed on a category label per word in the same HIT. We report below the percentage of aggregate pairwise agreements among the turkers for those categories. We report the results of the analysis by the majority class. Hence for those instances that are considered Hindi across the HITs, 98.1% of the times, some two turkers agreed on a Hindi label.

	h	**e**	**f**	**i**	**u**
h	167195	1875	370	535	426
e	2546	11800	229	47	215
f	578	253	3996	143	45
i	546	45	120	1467	29
u	442	212	40	32	99

Table 1: Confusion Matrix of the aggregate turkers' annotations for the different categories

All in all, the data had 193285 word instances, corresponding to 14658 word types, 88.16% word instances were considered Hindi by the majority of turkers, 7.67% instances were considered English by the majority of turkers, 2.59% words were considered Foreign Proper names, and 1.14% were considered Indian Proper names, finally 0.42% were considered Unknowns. The following statistics reflect the confusion on the majority label by aggregate pairs of turkers.

For majority class Hindi word instances (88.16% of the word instances):

Hindi- 98.1%
English- 1.1%
Foreign Proper Name- 0.22%
Indian Proper Name- 0.31%
Unknown- 0.25%

Hence, turkers agreed 98.1% of the time that the label for these 88.16% of the word instances are Hindi, however, some set of the turker pairs confused 1.1% of this Hindi data set as English, while 0.25% of the time pairs of turkers considered these Hindi words as Unknown.

For majority class English word instances (7.67% of the word instances):

Hindi- 17.16%
English- 79.53%
Foreign Proper Name- 1.54%
Indian Proper Name- .32%
Unknown- 1.45%

The turkers agreed 79.53% of the time that

the label for these 7.67% of the word instances are English, however, some set of the turker pairs confused 17.16% of this English data set as Hindi, 1.54% as Foreign Proper Name, 0.32% as Indian Proper Name and 1.45% of the time pairs of turkers considered these English words as Unknown

For majority class Foreign Proper Name word instances (2.59% of the word instances):

Hindi- 11.52%
English- 5.04%
Foreign Proper Name- 79.68%
Indian Proper Name- 2.85%
Unknown- .9%

The turkers agreed 79.68% of the time that the label for these 2.59% of the word instances are Foreign Proper Name, however, some set of the turker pairs confused 11.52% of this Foreign Proper Name data set as Hindi, 5.04% as English, 2.85% as Indian Proper Name and 0.9% of the time pairs of turkers considered these Foreign Proper Names as Unknown

For majority class Indian Proper Name word instances (1.14% of the word instances):

Hindi- 24.74%
English- 2.04%
Foreign Proper Name- 5.44%
Indian Proper Name- 66.47%
Unknown- 1.31%

For majority class Unknown word instances (0.42% of the word instances):

Hindi- 53.58%
English- 25.7%
Foreign Proper Name- 4.85%
Indian Proper Name- 3.88%
Unknown- 12%

The above statistics are the aggregated results, we note that the results for each of the 93 turkers taken individually, as compared to their respective co-turkers follow the same trend as the aggregated results. For example, if we compare an individual turker with co-turkers, majority of agreements are Hindi-Hindi, English-English and so on. Similarly, disagreements are also proportionate to above statistics.

A detailed token level analysis also showed similar trends. We analyzed a sample of 1304 tokens of which 1005 have a Hindi root and 245 are of English etymology. 27 tokens were Foreign Proper Names and 26 were Indian Proper Names. The turkers agreed 98.45% times that the tokens are Hindi over the total occurrences of sample Hindi root tokens. They agreed 79.41% times that the token is English over tokens with English root, 75.74% times agreed that the token is Foreign Proper Name for Foreign Proper Name tokens. The turkers agreed 74.58% times that the token is an Indian Proper Name for Indian Proper Name tokens. The turkers were observed to confuse Hindi tokens and Indian Proper Name tokens as 22.63% times, i.e. they mutually agreed that the token is Hindi when it was in fact an Indian Proper Name.

We further analyze the agreement on a complete sentence level, where turkers agreed on the annotation for every token in the sentence, we found only 57 such sentence annotations.

6 Analysis of Results

As depicted by the above statistics, the largest percentage of agreement was for the words marked in the Hindi category. There was about 98% agreement over such words which can be attributed to two reasons. Firstly, Hindi being the matrix language, a dominant part of words in the sentences were Hindi. Secondly, although there were very few instances where the turkers completely agreed on each word of a sentence, they had almost no confusion in identifying the Hindi words in a sentence.

For a word classified as English by a turker, the co-turkers agreed 80% times. However, about 17% co-turkers confused such words as Hindi. An obvious reason for such observation is the fact that some of the English words have blended so well with Hindi that even the native Hindi speakers are not able to recognize them as English words. For example, English words such as **cycle, car, train, plate, bread** have become part of Hindi lexicons and the native speakers unintelligibly consider these words as Hindi itself in their conversations. This shows the seamless mingling of English words in Hindi to such an extent that they are

indistinguishable as English words.

There was agreement for majority of Foreign and Indian Proper Names (79.68% and 66.47% respectively). The highest percentage of disagreements were observed when the co-turkers marked proper names as Hindi words. This may be attributed to the fact that capitalization does not exist in the Hindi script for proper names, and they might be misconstrued as other parts of speech. For example, **Pawan** could be a name and could be used as a noun meaning **air** as well. Similarly, **Anant** could be used as a name or an adjective meaning **with out an end**.

The Unknown category showed some interesting results. The agreement over Unknown category was much less among turkers. Instead, majority of co-turkers marked such words as Hindi words (53.58% times). After analysis, it was found that a major reason for this observation was because turkers were confused on morphologically mixed words. For example, plural of **company** after morphological adjustment becomes **companiyon** in Hinglish. A turker was not able to classify such words distinctly since it is half Hindi and half English from his point of view. Moreover, we believe that the fact that we have a default Hindi tag, could have contributed to the confusion. In our next iteration of annotation experiments, we will make sure to avoid a default tag.

If we consider the overall results, more than 90% of agreements were for Hindi words followed by English, Foreign Proper Name, Indian Proper Name and Unknown categories, in that order. This along with results for each of the individual categories shows that the turkers have high confidence while marking the Hindi words. English words, foreign proper names and Indian proper names also show good confidence with agreement over majority of them. Majority of disagreements in different categories were classified as Hindi which shows the tendency of the turkers to mark the word, about which they are confused, as Hindi itself, or simply leave it as default. Based on analysis of results above in conjunction with the inter-turker and gold agreements, it can be affirmed with high confidence that apart from the *Unknown* category words, the turkers converge on the correct category significantly above chance indicating the feasibility of the approach.

7 Conclusion and Future Directions

In this paper we presented an initial attempt at building a large scale repository of manually annotated LCS data for Hinglish. We believe we have established that crowd sourcing is a good method for inducing such annotations. In the near future we plan on annotating more data. We plan on adding a new category label of mixed morphology. Finally, we intend to perform the same annotation task for other language pairs.

References

Aravind Joshi. 1985. Processing of sentences with intrasential code switching. *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. Cambridge University Press, Cambridge, UK.

Carol Myers-Scotton. 1993. Common and Uncommon Ground: Social and Structural Factors in Codeswitching *Language in Society*, 22(4):475–503.

Rakesh M. Bhatt. 1997. Code-switching, constraints, and optimal grammars. *Lingua*, 102(4):223–251.

Rakesh M. Bhatt and Agnes Bolonyai. 2008. Code-switching and optimal grammars . *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 44(2):109–122.

Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Ng. 2008. Cheap and fast but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of the EMNLP 2008, Honolulu, Hawaii*, 254–263.

S Poplack. 2001. Code-switching (Linguistic). N. Smelser and P. Baltes (eds.) *International Encyclopedia of the Social and Behavioral Sciences*, 2062–2065.

Thamar Solorio and Yang Liu. 2008. Learning to Predict Code-Switching Points. *Proceedings of the EMNLP 2008, Honolulu, Hawaii*, 973–981.

Thamar Solorio and Yang Liu. 2008. Part-of-Speech Tagging for English-Spanish Code-Switched Text. *Proceedings of the EMNLP 2008, Honolulu, Hawaii*, 1051–1060.