Psycholinguistic Analysis of Code Mixing - SNLP Term Project

Avirup Saha, Soumi Das, Indrasekhar Sengupta, Ayan Chandra

Mentor: Jasabanta Patro

November 29, 2017

Introduction

Code switching or **code mixing** is a lexical phenomenon which refers to natural switching of words or phrases between more than one language. **Code borrowing** or **lexical borrowing** refers to the situation where words from one language (say L1) become part of the vocabulary of another language (say L2) due to widespread adoption. This occurs when

- the native language L2 lacks suitable words that convey the same senses appropriately
- foreign word usage dominates its equivalent native language due to wide popularity

Examples of code borrowing

- P(कॉलेज जाना) $>$ P(महाविद्यालय जाना)

Examples of code borrowing

- P(कॉलेज जाना) $>$ P(महाविद्यालय जाना)
- P(फ़िल्म देखना) $>$ P(चलचित्र देखना)

Examples of code borrowing

- P(कॉलेज जाना) $>$ P(महाविद्यालय जाना)
- P(फ़िल्म देखना) $>$ P(चलचित्र देखना)
- P(क्लास जाना) $>$ P(कक्षा जाना)

Examples of code mixing

- वह एक cool dude है

Examples of code mixing

- वह एक cool dude है
- restaurant में खाना

Examples of code mixing

- वह एक cool dude है
- restaurant में खाना
- यह train का time change हो गया हैं क्या?

Goals

In this project we aim to **characterize code borrowing and code mixing** from the psycholinguistic point of view. In particular, we wish to:

- Propose psycholinguistic based metrics for quantification and prediction of lexical borrowing from code mixing.

Goals

In this project we aim to **characterize code borrowing and code mixing** from the psycholinguistic point of view. In particular, we wish to:

- Propose psycholinguistic based metrics for quantification and prediction of lexical borrowing from code mixing.
- Compare our metrics with various social media based metrics.

Goals

In this project we aim to **characterize code borrowing and code mixing** from the psycholinguistic point of view. In particular, we wish to:

- Propose psycholinguistic based metrics for quantification and prediction of lexical borrowing from code mixing.
- Compare our metrics with various social media based metrics.
- Measure how efficiently our metrics improve the language tagging process as compared to baselines.

Design of Experiment

- We have used Psytoolkit, a free on-line psycholinguistic survey tool to perform empirical experiments to capture cognitive signals of lexical borrowing from the participants.

Design of Experiment

- We have used Psytoolkit, a free on-line psycholinguistic survey tool to perform empirical experiments to capture cognitive signals of lexical borrowing from the participants.
- We record user responses as well as reaction times.

Survey Configuration

- Participants were given a series of Hindi phrases containing translated and transliterated forms of 57 selected English words, along with some invalid phrases.

Survey Configuration

- Participants were given a series of Hindi phrases containing translated and transliterated forms of 57 selected English words, along with some invalid phrases.
- 1. Entire list of phrases has been broken down into three sets.
  2. Each phrase has been marked valid on pressing 'A' and invalid on pressing 'L'.
  3. Each phrase stays for 5 seconds before timeout.
  4. Participant took two minutes interval in between two consecutive sets of phrases.

Survey Configuration

- Participants were given a series of Hindi phrases containing translated and transliterated forms of 57 selected English words, along with some invalid phrases.
- 1. Entire list of phrases has been broken down into three sets.
  2. Each phrase has been marked valid on pressing 'A' and invalid on pressing 'L'.
  3. Each phrase stays for 5 seconds before timeout.
  4. Participant took two minutes interval in between two consecutive sets of phrases.
- >60 participants took part in the survey, out of which 47 participants completed the entire three set of surveys.

Survey Output

- For each participant, three files are generated with attributes **word, option indicating transliterated or translated or invalid, mark of the user and reaction time**.

Survey Output

- For each participant, three files are generated with attributes **word, option indicating transliterated or translated or invalid, mark of the user and reaction time**.
-
```
job Transliterated 3 5000
god Translated 3 5000
friend Transliterated 3 5000
development Transliterated 3 5000
gift Invalid 3 5000
```

Figure: Features of dataset

Survey Output

- For each participant, three files are generated with attributes **word, option indicating transliterated or translated or invalid, mark of the user and reaction time**.

- ```
  job Transliterated 3 5000
  god Translated 3 5000
  friend Transliterated 3 5000
  development Transliterated 3 5000
  gift Invalid 3 5000
  group Translated 3 5000
  ```

Figure: Features of dataset

- A reference file that holds the mapping between participant ID and the survey output files.
  ```

Defining Metrics

Table: Measures for Defining Metrics

	Transliteration	**Translation**
Valid	Valid Transliteration	Valid Translation
Invalid	Invalid Transliteration	Invalid Translation

Defining Metrics

- $$\text{Metric-1} = \frac{\text{Valid Transliteration}}{\text{Valid Translation}}$$

Defining Metrics

- $$\text{Metric-1} = \frac{\text{Valid Transliteration}}{\text{Valid Translation}}$$
- $$\text{Metric-2} = \frac{\text{Valid Transliteration}}{\text{Valid Translation} + \text{Invalid Transliteration}}$$

Defining Metrics

- $$\text{Metric-3} = \frac{\dfrac{\text{Valid Transliteration}}{\text{Average Reaction Time for Valid Transliteration}}}{\dfrac{\text{Valid Translation}}{\text{Average Reaction Time for Valid Translation}}}$$

Defining Metrics

- $$\text{Metric-4} = \frac{\frac{\text{Valid Transliteration}}{\text{Average Reaction Time for Valid Transliteration}}}{\frac{\text{Valid Translation}}{\text{Average Reaction Time for Valid Translation}} + \frac{\text{Invalid Transliteration}}{\text{Average Reaction Time for Invalid Transliteration}}}$$

Table: Words with Metric Values

Word	Metric-1	Metric-2	Metric-3	Metric-4
play	0.6190	0.4127	0.3137	1.472e-06
lyrics	0.7	0.5	0.4359	1.7699e-06
people	0.4222	0.2639	0.1334	1.2357e-06
uncle	1.3125	1.1666	1.6794	9.0603e-06
politics	0.5555	0.3906	0.3332	1.373e-06
review	0.8571	0.5882	0.7920	2.173e-06
parliament	0.7555	0.5965	0.5219	1.743e-06
house	0.5581	0.375	0.3107	1.459e-06
film	1.2286	1.1316	2.447	1.677e-06
god	0.5238	0.344	0.2265	9.757e-07

Table: Top Ranked Words as per Metric-1

Rank	Words
1	film
2	interview
3	college
4	uncle
5	body

Table: Top Ranked Words as per Metric-2

Rank	Words
1	film
2	college
3	uncle
4	interview
5	body

Table: Top Ranked Words as per Metric-3

Rank	Words
1	film
2	college
3	interview
4	uncle
5	body

Table: Top Ranked Words as per Metric-4

Rank	Words
1	film
2	college
3	uncle
4	interview
5	body

Top Ranked Words

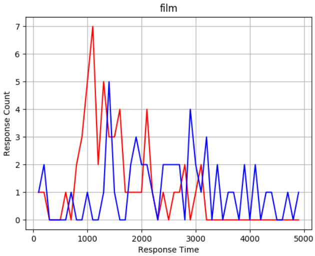

Figure: Word with rank-1 as per Metrics 1-4 (red: transliterated, blue: translated)

Top Ranked Words

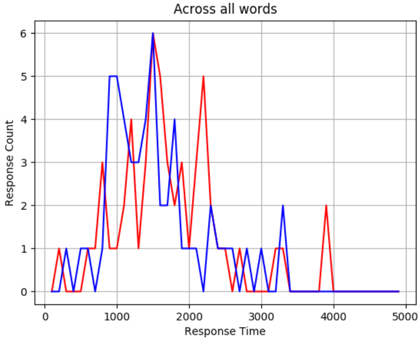

Table: Ground-Truth

Word	Survey_Rank	UUR	UTR	Log_ranking
blue	1	4	6	5
body	2	20	20	43
boy	3	2	2	8
car	4	23	23	30
class	5.5	28	29	19
college	5.5	14	14	15
cool	8	21	21	32
day	8	7	7	3
degree	8	16	16	34
development	10	15	15	41

Comparison with ground truth

Table: Spearman's Rank Correlation Coefficient of Metrics with Ground Truth (Survey Rank)

Metric	Correlation Coefficient
1	0.021173803109
2	0.0939040333899
3	0.0472114201021
4	0.0748703083899

Defining Metrics

- The reaction times for the responses "Valid Translation" and "Valid Transliteration" form a probability distribution over the range 0-5000 ms (due to timeout).

Defining Metrics

- The reaction times for the responses "Valid Translation" and "Valid Transliteration" form a probability distribution over the range 0-5000 ms (due to timeout).
- We divide the range (0, 5000) into N equal sized intervals

Defining Metrics

- The reaction times for the responses "Valid Translation" and "Valid Transliteration" form a probability distribution over the range 0-5000 ms (due to timeout).
- We divide the range (0, 5000) into N equal sized intervals
- We count the above responses in each interval and estimate corresponding probabilities by normalization by sum of all counts to obtain two probability vectors A_w and B_w for each word w.

Defining Metrics

- $$\text{Metric-5} = \|A_w - B_w\|_2$$

Defining Metrics

- $$\text{Metric-5} = \|A_w - B_w\|_2$$

- $$A_w(i) = P(Valid\ Translation \mid RT \in (t_i, t_{i+1})), i \in \{1, 2, \ldots, N\}$$

- $$B_w(i) = P(Valid\ Transliteration \mid RT \in (t_i, t_{i+1})), i \in \{1, 2, \ldots, N\}$$

Comparison of Metric-5 with ground truth

Table: SRCC of Metric-5 with Ground Truth (Survey_Rank) for varying interval size

interval size (ms)	SRCC (Metric-5)
50	0.0577821089436
100	0.163456571933
150	0.0621271160135
200	-0.0191634267035
250	0.00204280183134
300	0.0269455289182
350	0.0170233485945
400	0.00444228334752
450	0.0897860043013
500	0.0454604471038

Top Ranked Words as per Metric-5

Table: Top Ranked Words as per Metric-5 for interval size 100 ms

Rank	Words
1	well
2	boy
3	woman
4	question
5	friend

Conclusion and Further Work

We have conducted a study on psycholinguistic behavior of code mixing and defined metrics to indicate the likelihood of code borrowing. Future work includes:

- Pruning the existing metrics from experiment.

Conclusion and Further Work

We have conducted a study on psycholinguistic behavior of code mixing and defined metrics to indicate the likelihood of code borrowing. Future work includes:

- Pruning the existing metrics from experiment.
- Finding more appropriate metrics to capture the degree of code borrowing.

Conclusion and Further Work

We have conducted a study on psycholinguistic behavior of code mixing and defined metrics to indicate the likelihood of code borrowing. Future work includes:

- Pruning the existing metrics from experiment.
- Finding more appropriate metrics to capture the degree of code borrowing.
- Continuing the survey for user base of different age group and regions so that the degree of code borrowing can be indicated differently for each category.

References

- Jasabanta Patro, Bidisha Samanta, Saurabh Singh, Abhipsa Basu, Prithwish Mukherjee, Monojit Choudhury, Animesh Mukherjee, "All that is English may be Hindi: Enhancing language identification through automatic ranking of likeliness of word borrowing in social media", EMNLP 2017
- www.psytoolkit.org
- 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.