Borrowing Likeliness Ranking based on Relevance Factor

Rohan Agrawal
VIT University Chennai
India
rohan.agrawal2013@vit.ac.in

Rajalakshmi R
VIT University Chennai
India
rajalakshmi.r@vit.ac.in

ABSTRACT

Code mixing and code borrowing are the two important linguistic phenomena seen among the bilingual and multilingual speakers. The present scenario demands highly efficient methods to distinguish code borrowing from code mixing to quickly process the multilingual queries. To address this problem, we are in need of a new metric to rank different words according to their borrowing likeliness. In this paper, a new relevance based metric is proposed by applying statistics based approach. By performing various experiments on the social media data corpus containing more than 2.5 lakh tweets, the effectiveness of the proposed metric was studied.

KEYWORDS

Multi-lingual Information Retrieval, Code-Borrowing, Ranking Criteria, Relevance Factor.

1 INTRODUCTION

Code-Mixing [1] refers to mixing of two languages. In particular, when words and phrases of one language, say foreign language, is used while communicating in another language, say domain language, then code mixing is said to occur. This phenomenon is often seen in the communication among bilingual and multilingual speakers. In Code-Mixing people are subconsciously aware of the foreign origin of the Code Mixed word or the phrase. A similar linguistic phenomenon is called Code-Borrowing, where a word or a phrase from a foreign language is used as a part of the native vocabulary of the domain language. This phenomenon can be seen in the communication between monolingual people of a language where people often use borrowed words or phrases without being aware of the foreign origin of that word or the phrase. Moreover, code mixing and borrowing is a dynamic phenomenon, therefore, usage of words needs to be tracked regularly. To distinguish the native words from that of borrowed words, we need a method to find out the borrowing likeliness of foreign–origin words. In this paper, we made an attempt to define a new metric for ranking the words according to their borrowing likeliness.

This paper is organized as follows: Section 2 discusses the proposed statistics based approach in the development of new metric for finding the borrowing likeliness of candidate words. The experiments and results have been discussed in Section 3 followed by Conclusion in Section 4.

2 PROPOSED METHODOLOGY

In the proposed system, we have applied statistics based approach to rank the words by determining the borrowing likeliness. The following are the steps involved in development of this system viz., Data collection, Preprocessing, Tagging the tweets and computing the statistics based relevance factor.

In the data collection phase, we have used the social media data corpus provided as part of the data challenge. We imported the tweets using the tweet ids given in challenge dataset. In pre-processing phase, from the collected tweets and the information provided in the data sheet, we extracted all the unique English and Hindi words. Using the hints provided in the data challenge task, we tagged the tweets as English, Hindi, Code Mixed English (CME), Code Mixed Hindi (CMH) and Code-Mixed Equal, (CMEQ).

Among the various tags associated with the tweets, as we are more interested in finding out the borrowing likeliness of English words in Hindi tweets, we have restricted ourselves to two categories only viz., CME and CMH for further processing.

Statistics based Ranking using Relevance Factor:

The supervised techniques such as Chi-square method [3] has been used in text classification for finding the dependency of the terms with various categories. The problem of finding the borrowed likeliness of a word can be viewed as a binary classification problem. Here, we need to find out the most commonly used English terms in non-English queries. So we have considered the CMH and CME tweets for developing a metric that defines the borrowing likeliness of English words in Hindi tweets. In the proposed methodology, if more than 50% of the tweet words are in English, we tagged those tweets as CME tweets. If more than 50% of the words are in Hindi, then we tagged those tweets as CMH tweet.

We have the collection of unique English words and Hindi words in the corpus. We form a 2x2 contingency table called CHI table for all the candidate words (230 words) provided in the data challenge as shown in Table 1. In this table, we present the observed frequency (O) of every word (w) in English tweets and non-English tweets. Let S represent this collection of 230 words. For each word w in S, we form the CHI table by determining observed frequencies. Here ‘a’ denotes the number of tweets in CME set that contain the word w; ‘b’ denotes number of tweets in CME set that do not contain the word w, ‘c’ denotes number of tweets in CMH set that contain the word w and ‘d’ denotes number of tweets in CMH set that do not contain the word w. These values are referred as observed frequencies (*O (i, j)*), if their cell position is (i,j).

	Word	⌐Word
CME	a	b
⌐ CME	c	d

Table 1: CHI ^{2} table

From the observed frequencies, we can calculate the expected frequency as shown below.

$$E(i,j) = \frac{column_i\ total \times row_j\ total}{a+b+c+d}$$

As we are interested to rank the English words according to their borrowing likeliness in Hindi, we focus on the value of 'c' alone ie. O(2,1) . For our ranking purposes, the expected value for this cell position (2,1) denoted by E(2,1) is calculated as shown below:

$$E(2,1) = \mathrm{E(c)} = \frac{(a+c)\times(c+d)}{(a+b+c+d)}$$

We now calculate the relevance of this term. Relevance for this term would be given by:

$$R = \frac{c}{E(c)}$$

This relevance value can be used as a metric for ranking the words on the basis of their borrowing likeliness. The higher the value of R for a word, the more likely it is to be borrowed. In this way, we can find out the most commonly used English terms that are borrowed and used Hindi. This can be helpful in retrieving the monolingual documents even though the query may be represented in multi-lingual form.

3 EXPERIMENTS AND RESULTS

The experiment was performed on a corpus of 2,58,757 tweets as provided in the data challenge. Python 2.7 programming language was used for the programming purposes. *Tweepy* library was used to import the tweets from Twitter. Among the 2,58,757 tweets, we were able to successfully collect 1,96,494 tweets. Based on the language tagging provided in the given challenge, the words in the tweets were tagged as English or Hindi. The tweets were then categorized as CME and CMH based on the above mentioned classification metric. The observed and expected frequencies for each of the 230 words were then calculated after that. Table 2 shows the results of system for a few sample words:

Word	a	b	c	d	Relevance
Bus	49	167099	30	19573	3.617
Uncle	25	167123	10	19593	2.721
Match	5107	162041	804	18799	1.295
Season	124	167024	5	19598	0.369
Person	503	166645	2	19601	0.037

Table 2: Sample results for the proposed relevance factor

It can be derived from the relevance values that the borrowedness index of *Bus* is higher than *Person* i.e. *Bus* is more likely to be code borrowed as compared to *Person*. Similarly, a table containing the relevance values of all the terms arranged in descending order would represent the ranking of the words in terms of borrowing likeliness.

Spearman's Rank-Order correlation calculation:

The results obtained using the proposed metric was then compared to the ground truth ranking using the Spearman's Rank-Order correlation.

The correlation factor was calculated using:

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$$

where d_i = difference in ranks calculated by the proposed metric and the ground truth reality. n = number of cases. Spearman's Rank-Order correlation for the 70 word test data was calculated and it is 0.3322.

4 CONCLUSIONS

In this paper, a new relevance based metric was proposed by applying statistics based approach. The effectiveness of this metric has been studied using other metrics such as CHI square value and it is found that, the proposed metric is better than the existing methods in ranking the words based on the borrowing likeliness. Combining linguistic based features with the proposed method will be explored in the future work.

REFERENCES

[1] Bali et al. (2014): Kalika Bali Jatin, Sharma , Monojit Choudhury, and Yogarshi Vyas. "'I am borrowing ya mixing?" An Analysis of English-Hindi Code Mixing in Facebook." EMNLP 2014 (2014): 116.

[2] Gella et al, (2013): Spandana Gella, Jatin Sharma and Kalika Bali. Query word labeling and Back Transliteration for Indian Languages: Shared task system description In Proceedings of the Fifth Workshop on Forum for Information Retrieval (FIRE 2013). New Delhi, India

[3] Oakes et al: Michael Oakes, Robert Gaizauskas, Helen Fowkes. A Method Based on the Chi-Square Test for Document Classification In *SIGIR'01,* September 9-12, New Orleans, Louisiana, USA. ACM 1-58113-331-6/01/0009

[4] "Spearman's Rank-Order Correlation - A Guide To When To Use It, What It Does And What The Assumptions Are.". Statistics.laerd.com.