Code-Borrowedness of English words in Hindi Language

Ram Mohan
R&D Department
Flytxt, India
ram.mohan@flytxt.com

Muhammad Arif
R&D Department
Flytxt, India
muhammad.arif@flytxt.com

Jobin Wilson
R&D Department
Flytxt, India
jobin.wilson@flytxt.com

nce format:
Ram Mohan, Muhammad Arif, and Jobin Wilson. 2016. Code-Borrowedness of English words in Hindi Language. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 2 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The goal of IKDD-CODS 2017 data challenge is to develop a metric to rank a set of candidate words according to the borrowing likeliness.

1.1 Evaluation Metric

To determine the ground truth a survey would be conducted where a user has to choose a Hindi statement with English word or a Hindi word used in the statement. The words are ranked based on the sum('English word preferred') - sum('Hindi word preferred')

1.2 Data Provided

i. To come up with a metric we are shared with twitter data along with words tagged as Hindi(HI) or English(EN) or NE(Named Entity) or OTH(OTHER) ii. Survey Responses for a subset of 12 words.

2 METRIC PROPOSED

As can be seen in block diagrams 1 and 2 the metric proposed involves identifying the features and learning the weights for the features which are elaborated in the following section.

2.1 Pre-processing and Feature Extraction

2.1.1 Tag Tweets. Tweets were tagged as HI or EN or CMH or CME or CMEQ. For tagging the tweets only EN(English) and HI(Hindi) words were considered (ie, OTH and NE(named entity) were not included for counting of total words for a tweet).

Table 1: Tweet Tagging Rules

Tweet Tag	Rule
EN	EN/(EN+HI)¿.9
HI	HI/(EN+HI)¿.9
CME	EN/(EN+HI)¿.5
CMH	HI/(EN+HI)¿.5
CMEQ	EN/(EN+HI)==.5

Conference'17, Washington, DC, USA
2016. 978-x-xxxx-xxxx-x/YY/MM...$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

2.1.2 Hash Tags. It was noticed that there were a lot of English tweets when compared to Hindi and CMH tweets, which was adding bias/noise of its own, to reduce it we filtered only those tweets which were either Hindi or CMH or English/CME/CMEQ tweet which have the hash tag of any Hindi/CMH tweet. It can also be understood over here as, when we conduct the survey for determining ground truth each of the keyword involved has an equal prior, where as if there are too many English tweets which are irrelevant then they introduce a noise.

2.1.3 Stemming and feature statistics. It was identified that words such as film/films and they need to be correctly processed, hence stemming was performed which gave us a cleaner statistics for the keywords involved.

Natural Language Toolkit(NLTK) was used for pre-processing. After removing special characters, tweets were tokenized into words, converted to lowercase, filtered out stopwords and individual words were stemmed to its root form.

Feature list mentioned in the problem statement were used as is except that they were extracted from tweets based on hinid HashTag HHTU$_h i - Isthenumberofuserswhohaveusedthekeywordintheir HinditweetsHHT$
$Isthenumberofuserswhohaveusedthekeywordintheir EnglishtweetsHHTU_c mh -$
$Isthenumberofuserswhohaveusedthekeywordintheir CMHtweetsHHTU_c me -$
$Isthenumberofuserswhohaveusedthekeywordintheir CMEtweetsHHTU_c meq -$
$Isthenumberofuserswhohaveusedthekeywordintheir CMEQtweetsHHTT_h i -$
$Isthenumberofhinditweetscontainingthekeyword HHTT_e n - Isthenumberofengli$
$Isthenumberofcmhtweetscontainingthekeyword HHTT_c me - Isthenumberofcmet$
$Isthenumberofcmeqtweetscontainingthekeyword$

2.2 Metric

Table 2: Spearmanr Correlation Coefficient for Algorithms

Algo	3 fold CV	12wor S
Ordinal Regression	-.03	0.97
Linear Regression	.67	.91
Non-Linear Regression	.73	.94
Neural Network		
HandCrafted	.7915	.7915

Table 2 tabulates the different algorithms, their cross-validated Spearmanr Correlation Coefficients and their 12 word trained, Spearmanr Correlation Coefficients. As there were very few data points available and we were not able to add data points our choice of model is Handcrafted weights model. The same has been uploaded

2.2.1 Handcrafted – Final Metric. The metric used for ranking was HindiTaggedUserRatio(w)=(HHTU$_h i + HHTU_c mh)/(HHTU_e n)$

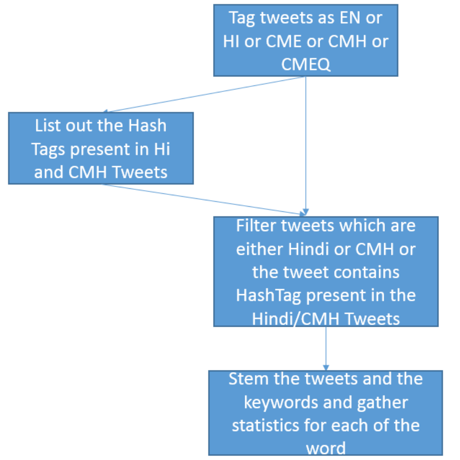

Figure 1: Preprocessing and Feature Extraction

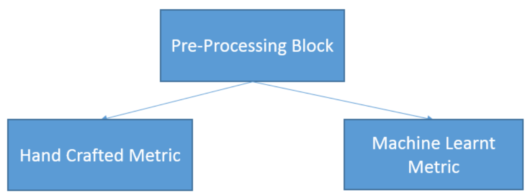

Figure 2: Learning weights of the Metric

HindiTaggedTweetRatio(w)=(HHTT$_h i + HHTT_c mh)/(HHTT_e n)$
FinalScore(w)=(HindiTaggedUserRatio(w)+HindiTaggedTweetRatio(w))/2

2.2.2 Machine Learning based– For Discussions only. For any supervised Machine Learning algorithm to be successfull we require sufficient data, hence we tried to devise multiple ways to add more keywords with ranks but we were confident with the extra data added. But still we will list down the models that were tried and were found to be promising on the 12 samples that were available.

1. Ordinal Regresssion/Learning to Rank : As part of this the features were made to learn the ordering 2. Linear Regression: As part of this weights for features were learnt against the normalized ground truth as y. 3. Non-Linear Regression: Kernalized-Ridge and SVR were modelled with ground truth as the y. 4. Neural Network: A neural network was trained for the model $y = (w_1'X + a)/(w_2'X + b)$.

3 CONCLUSIONS

In a limited data domain, The handcrafted metric along with proper preprocessing performed better.