Bhavya Pant
August 2019

Word Association in Non-literal Hindi Data

A study of Bollywood Song Lyrics

- *Abstract*

The central focus of this research project was creating a corpus of Bollywood song lyrics and performing word-association experiments on the resulting data, with the intention of comparing these results to those from similar tests performed on a more traditional corpus of literal Hindi data (in this case the *BBC Hindi News Corpus*.)

- *Background*

Indian cinema is the largest, most prolific film industry in the world, producing an average of 1600 films each year in Hindi, Marathi, Tamil and various other regional languages**.[1]** Of these, Bollywood films written primarily in Hindi, tend to make up the largest share. In these movies, original music is a characteristic motif and Bollywood soundtracks are hugely popular, accounting for nearly 80% of the music industry's revenue. In fact, T-Series, India's largest music record label has the most-viewed channel on Youtube with 80 billion views and over 109 million subscribers. Bollywood songs - composed primarily in Hindi (although they often borrow vocabulary from Urdu, Punjabi and increasingly English) - are hence a rich source of non-literal Hindi data.

- *Methodology*

We created the Bollywood lyrics corpus by scraping the web for song lyrics written in Devanagari script. www.hinditracks.in **[2]** proved to be a useful source that regularly updates its lyrics database and classifies its songs by movies as well as by artists. After preprocessing the data i.e. cleaning it up and removing any punctuation, numbers or other Roman text (including English words and Hindi lyrics transliterated into Roman script,) we ended up with a total of 1469 songs comprising of about 250,000 words.

These songs were then annotated by part-of-speech using the ISC POS Tagger **[3].** Then, the most frequently occurring tag bigrams (and their corresponding word pairs) from the corpus were examined. Of these, the top 50 (Adjective, Noun) i.e.(JJ, NN) and (Noun, Noun) i.e. (NN, NN) bigrams were isolated and subjected to word association and similarity tests.

The word association metric used in this experiment was the Positive Pointwise Mutual Information (PPMI) score. PPMI compares the probability of observing x and y together (the joint probability) with the probabilities of observing x and y independently(chance).

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

It follows that if there is a genuine association between x and y,then the joint probability P(x,y) will be much larger than chance P(x) P(y), and consequently I(x,y) >> 0. Whereas, if x and y are in complementary distribution, then P(x,y) will be much less than P(x) P(y), forcing I(x,y) << 0.**[4]**

The similarity scores for the words in the top bigrams were calculated using the Extended Lesk formula [5] whereby the glosses of the two words (as well as their respective hypernyms) were traversed for overlapping segments. The longer the overlapping the overlapping segment, the larger its contribution to the overlapping score.

$$\begin{aligned} similarity(A,B) = {} & overlap(gloss(A), gloss(B)) + \\ & overlap(gloss(hyper(A)), gloss(hyper(B))) + \\ & overlap(gloss(A), gloss(hyper(B))) + overlap(gloss(hyper(A)),gloss(B)) \end{aligned}$$

Word glosses were derived using the *pyiwn* - a Python based API to access the Hindi WordNet.**[5]**

As a means of comparison, similar annotation and analysis was then performed on the BBC Hindi News Corpus. The results of these two experiments were then examined comparatively. .

- *Initial Hypothesis*

Our initial hypothesis was that song lyrics data deals much more with metaphors and symbolism than data found in the BBC Hindi news corpus and should therefore mirror conversational, colloquial Hindi better than the latter. By calculating similarity scores of frequently occurring bigrams across the two corpora, we expect to find less similar or more distant bigrams in the bollywood lyrics corpus as compared to the News corpus. This should hopefully reveal some interesting insights into colloquial Hindi that are otherwise absent in similar analyses of literal corpora.

- *Observations*

Some initial preprocessing revealed that our Bollywood song lyrics corpus consisted of 21,850 unique tokens and averaged at about 180 tokens per song. Noun(NN) was unsurprisingly the most frequently occurring part-of-speech(POS) tag, whereas (Noun, Preposition) was the most frequent tag bigram.

Similarly for the much larger news corpus (containing about 30 million words,) NN was also the most frequently occurring POS tag, while (Noun, Postposition) was the most frequent tag bigram. Aside from that, there was little overlap between the top bigrams across the two corpora. Bigrams like (VM, PRP) i.e. (Verb, Preposition) present abundantly in the lyrics corpus were nearly obsolete in the news corpus. This is because the lyrics corpus saw phrase structures like that seen in “खुश **रहे तू** सदा, ये दुआ **है मेरी**” which if present in the news corpus would likely take the form of “**मेरी** ये दुआ **है** कि **तू** सदा खुश **रहे** ”

Whereas the top (NN, NN) bigrams for the news corpus contained quite a few multi-word expressions, such as “भूख हड़ताल” *(hunger strike)* and “सूचना पत्र” *(newsletter,)* those for the lyrics corpus were composed almost entirely of repeated words, such as “दिल दिल” and “प्यार प्यार.” This resulted in the similarity scores for the (NN, NN) bigrams in the lyrics corpus being much higher than for those in the news corpus.

- *Challenges*

Even after initially cleaning the corpus, it still contained some non-Hindi words. However, these had been transliterated into Devanagari and had resultantly slipped through the RegEx filter. These muddied up the data and resulted in some inaccuracies in the tagging and subsequent analyses.

The POS tagger gave out some additional erroneous data, for instance occasionally tagging 'तू' as an adjective. A future project could involve re-training the tagger model by manually re-tagging the inaccuracies.

Some tokens slipped through the Hindi WordNet API. Of the top 100 words present in the lyrics corpus, around 30 could not be found in the WordNet, owing largely due to their case markings. This in turn impacted the word similarity metrics.

Lastly, the small size of our Bollywood lyrics corpus (especially when juxtaposed with the BBC news corpus) was a major drawback. A larger corpus would have been able to better overcome intermittent losses in data and reveal further insights

- Resources
    - www.hinditracks.in is an online database of Bollywood song lyrics written in Devanagari script.
    - The ISCNLP POS Tagger for Indian Languages was used to perform part-of-speech tagging on the scraped data.
    - BBC Hindi News Corpus
    - IIT Bombay Python WordNet for Indian Languages
    - Behl, Aseem and Monojit Choudhury. *“A Corpus Linguistic Study of Bollywood Song Lyrics in the Framework of Complex Network Theory.”* (2011).
    - Understanding Bollywood lyrics -
    - PPMI paper (hanks churcoli)
    - Jarufsky Martin