Abstract

Finding semantic similarity touches wide range of problems in real life scenarios. It may be used to summarize the news articles, to find similarities in the reviews of some product so as to get pros and cons of a product, etc. For doing semantic similarity, the problem with Indian language is that there is no proper parser available available for the Indian languages. Hence we proceed through WordVector implementation.

Motivation

- No proper parser available for Indian language.
- Local language user makes up to large user base on the internet.

Dataset

Hindi news dataset of about 300 news articles

Project Goals/Objectives

- Collect and create dataset in Hindi language.
- Word level similarity using WordVector.
- Sentence level similarity using word level similarity.
- Semantic Vector for sentences.
- Being able to merge news articles using the above sentence similarity

Methods/Process

The approach is similar to Yuhua Li et al[2] for “Sentence Similarity Based on Semantic Nets and Corpus Statistics”.

Get hierarchical knowledge base by using Word2Vec and then getting a raw word similarity based on common synsets. We take the joint word set of the two sentences that we have to compare then we make an n-dimensional vector each for sentence T1 and T2 where n = #words in T1 + #words in T2 followed by calculating weights of each dimension as:-

\[
S(T_1, T_2) = \delta S_s + (1 - \delta)S_r
\]

Once we have the results, recognize the threshold to classify similar and dissimilar by ROC curve. The binary classifier can then be used to merge the news articles based on what are similar and dissimilar by ROC curve.

Conclusions/Recommendations

An approach was given to calculate sentence similarity score with the help of Word2Vec. Sentences having similar words tend to give more score irrespective of the overall sentence meaning, hence considered the impact of word order on sentence meaning. However, still without proper word disambiguation, similarity score was much higher then human score in some cases like in example 5 and vice-versa as in example 6. However in real life scenarios, sentences having many similar words tend to convey same meaning most of the time.

References

- Poster Template:- https://goo.gl/dd5j74
- Userbase information:- http://goo.gl/vqnhNU
- Userbase information:- http://goo.gl/6nTxEP
- List of hindi stop words:- https://goo.gl/a0gRzd