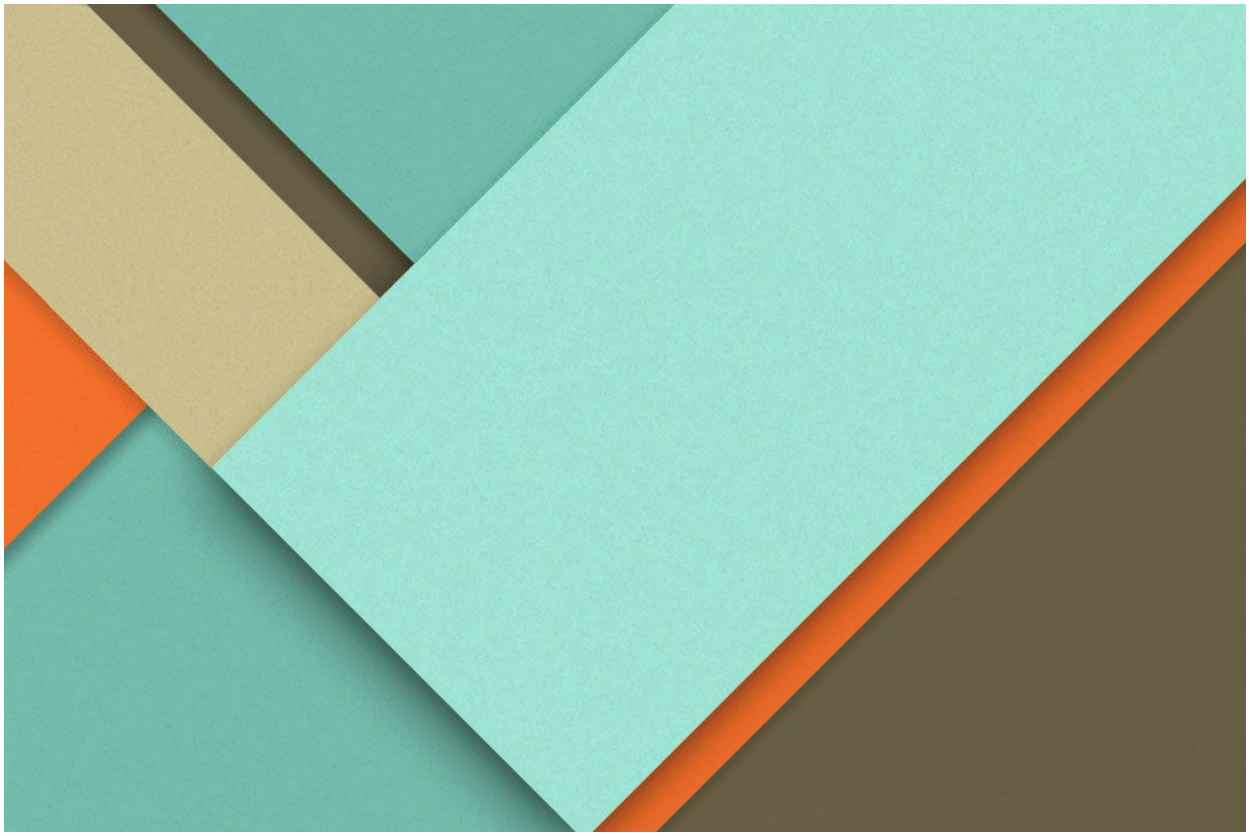

Semantic Similarity using WordVector

09.10.2015

Rakshit Sharma
Dept. of Computer Science
IIT Kanpur
Kanpur-208016

Advisor:- Prof Amitabha Mukerjee
Dept. of Computer Science
IIT Kanpur
Kanpur-208016

Introduction & Motivation

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.

Related Works

A lot of work has been done and is being done on the semantic similarity. This has been a consistent topic since last 4 years in SemEval events[1]. Each of the papers submitted in these tasks have its own advantages and disadvantages. But most of them work on usage of parsers to preprocess the input, here we propose to use a method that does not use any parser and is independent of any specific language structure.

Approach

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

So, we'll be getting hierarchical knowledge base by using WordNet/Word2Vec/Glove and then getting a raw word similarity based on common synsets. Followed by training the data on a corpus to get actual semantic word similarity.
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 T_1 and T_2 where n = #words in T_1 + #words in T_2 followed by calculating weights of each dimension as:-

$$\left\{ S1[i] = \begin{matrix} 1 \\ similarity\ score \end{matrix} \quad \begin{matrix} \forall w[i]\ in\ the\ joint\ word\ set \in T1 \\ \forall w[i]\ in\ the\ joint\ word\ set \notin T1 \end{matrix} \right\}$$

$$\left\{ S2[i] = \begin{matrix} 1 \\ similarity\ score \end{matrix} \quad \begin{matrix} \forall w[i]\ in\ the\ joint\ word\ set \in T2 \\ \forall w[i]\ in\ the\ joint\ word\ set \notin T2 \end{matrix} \right\}$$

where S1 and S2 are similarity score vectors for T_1 and T_2. And then we calculate overall sentence similarity as:-

$$S = \frac{S1.S2}{||S1|| + ||S2||}$$

References

[1] SemEval 2015(task 1):- http://alt.qcri.org/semeval2015/
SemEval 2014(task 3):- http://alt.qcri.org/semeval2014/
SemEval 2013(task 6):- https://www.cs.york.ac.uk/semeval-2013/
SemEval 2012(task 6):- https://www.cs.york.ac.uk/semeval-2012/

[2] Yuhua Li et al "Sentence Similarity Based on Semantic Nets and Corpus Statistics" (AUGUST 2006) (IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 8)
http://ants.iis.sinica.edu.tw/3BkMJ9lTeWXTSrrvNoKNFDxRm3zFwRR/55/Sentence%20Similarity%20Based%20on%20Semantic%20Nets%20and%20corpus%20statistics.pdf