Name – KANAD DAS

Roll No. – B18733

University – RAMAKRISHNA MISSION VIVEKANANDA EDUCATIONAL & RESEARCH INSTITUTE

Program Course – BIG DATA ANALYTICS (BDA)

Program Release Date – April 04, 2019

Date of Submission – April 28, 2019

Opinion Mining using WordNet

Introduction

Aim: This is an assignment on the problem of text mining. We are given a dataset of opinions of a group of people regarding the question "*What qualities do you think are necessary to be a prime minister of India?*" The task is to find the significant qualities from the opinions using Wordnet, an English lexical database and analyse the performance of our method by comparing our results with the human coding i.e., significant qualities found by some experts.

What is Wordnet?

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

WordNet groups words together based on their meanings. However, there are some important characteristics. First, it interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found near one another in the network are semantically disambiguated. Second, it labels the semantic relations among words.[1]

Dataset: The dataset contains the opinions of 38 different people on the following question:

> "*What qualities do you think are necessary to be a prime minister of India?*"

The opinions are mainly short texts which state the qualities of an ideal prime minister according to the respective people.

Methodology

The steps involved in finding the significant qualities from the given dataset are as follows:

1. We form the temporary vocabulary of the dataset by considering only the nouns and adjectives from the text. We ignore the other words since the objective is to find the qualities of the prime minister. We then find the frequency of each term in the temporary vocabulary and arrange the terms in order of decreasing frequency.
2. We find the synsets with respect to nouns and adjective only, for each word using Wordnet and store them in a list. We then find the derivationally related forms of each synset and then find the synsets of those forms and store them in another list. Finally, we add these two lists and form the main vocabulary containing the synsets of both the original terms and the derivationally related forms.
3. Next, we find similar terms from the main vocabulary by using Wu-Palmer's Similarity and merge those terms to form a cluster, for which the similarity score is greater than a given threshold value (**= 0.9**).
4. We ignore the empty clusters and assign the remaining clusters to the original terms of the text if the term is present in the cluster. Here, we combine the terms which have the same clusters.
5. Finally, we obtain the names corresponding to each synset present in the clusters of each term. We then recalculate the frequency of each cluster and arrange them in decreasing order of frequency.

Wu-Palmer's Similarity: Let $C1$ and $C2$ be two concepts in the taxonomy. This similarity measure considers the position of *C1* and *C2* to the position of the most specific common concept *C*. Several parents can be shared by *C1* and *C2* by multiple paths. The most specific common concept *is the* closest common ancestor *C* (the common parent related with the minimum number of IS-A links with concepts $C1$ and $C2$).

$$Sim_{wup}(C1, C2) = \frac{2N}{N1 + N2 + 2N}$$

where *N1* and *N2* are the distance (number of IS-A links) that separates, respectively, the concept *C1* and *C2* from the specific common concept and *N* is the distance which separates the closest common ancestor of *C1* and *C2* from the root node. [2]

Path Similarity: This measure considers only the length between *C1* and *C2*. It measures the similarity of concepts, depending on how close they are to each other in the taxonomy. [3] It is based on two observations. One is that the behaviour of conceptual distance resembles that of a metric. The other is that the conceptual distance between two nodes is proportional to the number of edges separating the two nodes in the hierarchy. [4]

Results and Analysis

Using WordNet, we found a total of **1083** clusters of synsets of both the original words in the vocabulary as well as their derivational forms. After removing the empty clusters, the number decreased to **1076** clusters. After assigning these clusters to the original terms in the vocabulary based on their presence, the number of clusters scaled down to **94**. This was achieved with the help of python dictionaries where the original terms became the keys and the clusters their corresponding values. The remaining clusters were discarded since absence of the original terms in them rendered them useless to completing our task. In these pairs of terms and clusters, there were few terms with same clusters which were then combined to a single cluster. Hence the final result contains **92** clusters, corresponding to the original terms in the vocabulary.

Out of these 84 clusters, if we consider only those clusters whose frequency are greater than 2, then we can see that these clusters contain **5-6** significant qualities required in a prime minister as analysed by the experts, and thus we can say that the clusters obtained by this similarity score are good as compared to the human encoding.

A **significant finding** from these top clusters is that, apart from finding the qualities as found by human encoding, the clusters also found some other significant qualities required in a prime minister, e.g.- **morality**, **integrity**, **experience**, **dedication**, etc.

The above results are obtained by using Wu-Palmer Similarity score. For results obtained by Path Similarity score, see Appendix.

Conclusion

Here the framework or the methodology solely depends on WordNet.

The main advantage of WordNet is that being somewhat an online thesaurus, it finds words which are similar on a lexical basis. Thus, it provides words which are similar in sense to a greater extent.

Though WordNet is a useful tool, it has quite a few limitations when it comes to natural language processing.

1. Even though cast in a general cognitive framework, Wordnet still assumes a lexicalist/projectionist view of lexical properties, meaning it assumes most if not all lexical properties (syntactic, semantic properties) can be encoded at the lexical level, regardless of context.
2. Wordnet does not provide a clear distinction criterion between atomic and non-atomic lexical units. In other words, it is hard to distinguish between "pure" simple words and multi-word units, such as collocations, verbal phrases, frozen expressions, idioms, etc.
3. Frequency data are not available, so it's hard to decide whether "Proud" is more common than "Beaming", for example.

References

[1] Princeton University "About WordNet." WordNet. Princeton University. 2010.

[2] Z. Wu and M. Palmer. 1994. Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics (ACL'94)*, pages 133-138, Las Cruces, New Mexico.

[3] R. Rada, H. Mili, E. Bicknell and M. Blettner, “Development and Application of a Metric on Semantic Nets”, IEEE Transactions on Systems, Man and Cybernetics, vol. 19, Issue 1, (1989) January-February, pp. 17 - 30.

[4] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis and E. E. Milios, “Semantic similarity methods in WordNet and their application to information retrieval on the web”, Proceedings of the 7th annual ACM international workshop on Web information and data management, (2005) October 31- November 05, Bremen, Germany.

Appendix

Results obtained by using Path Similarity score:

We found a total of **1083** clusters of synsets of both the original words in the vocabulary as well as their derivational forms. There are no empty clusters. After assigning these clusters to the original terms in the vocabulary based on their presence, the number of clusters scaled down to **95**. This was achieved with the help of python dictionaries where the original terms became the keys and the clusters their corresponding values. The remaining clusters were discarded since absence of the original terms in them rendered them useless to completing our task.

Out of these 95 clusters, if we consider only those clusters whose frequency are greater than 2, then we can see that these clusters contains **2-3** significant qualities required in a prime minister as analysed by the experts, and thus we can say that the clusters obtained by this similarity score are not good compared to the human encoding.