Measuring Text-Based Semantics Relatedness Using WordNet

Madiha Khan, Sidrah Ramzan, Seemab Khan, Shahzad Hassan, Kamran Saeed

Abstract**—**Measuring semantic similarity between texts is calculating semantic relatedness between texts using various techniques. Our web application (Measuring Relatedness of Concepts-MRC) allows user to input two text corpuses and get semantic similarity percentage between both using WordNet. Our application goes through five stages for the computation of semantic relatedness. Those stages are: Preprocessing (extracts keywords from content), Feature Extraction (classification of words into Parts-of-Speech), Synonyms Extraction (retrieves synonyms against each keyword), Measuring Similarity (using keywords and synonyms, similarity is measured) and Visualization (graphical representation of similarity measure). Hence the user can measure similarity on basis of features as well. The end result is a percentage score and the word(s) which form the basis of similarity between both texts with use of different tools on same platform. In future work we look forward for a Web as a live corpus application that provides a simpler and user friendly tool to compare documents and extract useful information.

Keywords**—**GraphViz representation, semantic relatedness, similarity measurement, WordNet similarity.

I. Introduction

THIS paper describes how semantic similarity can be employed to extract useful common information between two documents. The document discusses how we can understand the extent of similarity between two documents through simpler a mechanism. So, let us first find what semantic similarity itself is exactly. Semantic similarity refers to the actual meaning of words in the sentence rather than characters. This similarity is based on ‘is-a’ and ‘has-a’ relationship between the words. Semantic similarity plays significant role in various fields like text mining, in data mining high quality information can be extract by applying semantics similarity techniques.

Up till now, researchers in [2], [3] and [7] have discovered different techniques in order to find semantic similarity. One of them, for example, is natural language processing which allows us to interpret the extent of similarity between documents. In the following research paper, we have taken up a software specified for finding semantic similarity, which is WordNet. This tool helps us find the similarity between two documents through the synonyms of actual words. Through use of synonyms, it represents explicit relationship between two words instead of implicit relation build on occurrences and usage.

The remainder of the paper follow the flow in such manner: Session II is comprised of methodology for calculating the semantic similarity between texts and their visualization. In Section III, we have concluded the current research and suggested developments to be made in future work. In Session IV we have acknowledged people who guided us and helped us while conducting the research. Session V is followed by the related work done in the respective field.

II. Methodology

The class diagram of our project is shown in Fig. 1 which describes all the methods and properties used for calculating similarity. The workflow of our system in Fig. 2 (attached at the end) shows the connection between different attributes.

The extracted content is processed through stop word tool which in turn passes keywords to porter class. Porter class stems keywords to their root form and passes to POS library. POS library splits keywords into parts of speech and forwards feature wise extracted keywords.

NHunspell takes keywords as input and retrieves synonyms against each keyword. These synonyms are stored and a separate dictionary is developed in database.

At last, semantics db takes keywords and synonyms as input and assigns each word a weigh. These words are placed in arrays and array of two documents are compared. Semantics db calculates the correlation and gives a score of similarity between both documents.

A. Preprocessing

In pre-processing, firstly stop words from word string are removed and then keywords are saved in database. In this process, array of content is compared word by word against stop word dictionary and noise is removed. We obtain a space separated list of keywords from it.

In next step, all the keywords are stemmed to root form so as to be able to get their synonyms. This is done by Porter’s stemmer which allows us to stem words so as to extract synonyms of each word.

Porter’s stemmer is a well-built stemming class which allows user to stem keywords to their root form. Synonyms can only be retrieved when keywords are in their root form because many words are unrecognizable for WordNet in their other

Madiha Khan, MS (Computer Engineering), is with the Bahria University, Pakistan (e-mail: madiha1940@gmail.com).

Sidrah Ramzan, continued her career as business development officer at Irtifa Group of Companies, Pakistan (email: sidrahramzan736@gmail.com)

Seemab Khan completed BS (Computer Science) from Air University, Pakistan (email: mmabi1097@gmail.com).

Dr.Shahzad Hassan, professor, is with the Computer Engineering Department, Bahria University Islamabad, Pakistan (e-mail: shassan.buic@bahria.edu.pk).

Kamran Saeed, research assistant, is with the Bahria University, Pakistan (e-mail: kamransaeed32@mts.ceme.edu.pk).

forms. Therefore, stemming process is essentially important to provide the input of keywords in their first simple form.

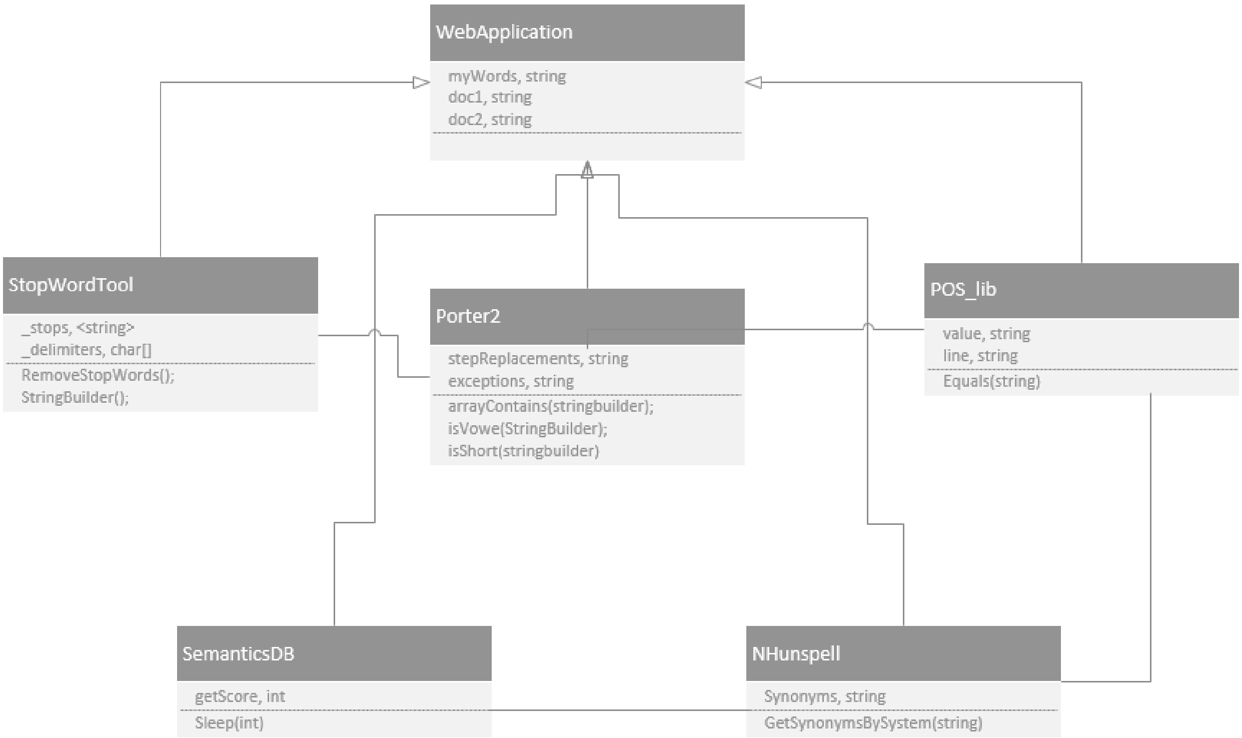

Fig. 1 Class Diagram

B. *Feature Extraction*

In feature extraction, the keywords and synonyms are taken as input arrays and are compared in loop. These words are compared against Parts of Speech (POS) syllables. The feature extractor then takes array of words as input and checks if word is a part of speech, for example, verb. If it is a verb, it is saved to verb database column. If it is not a verb then it is compared with the next syllable. Same method is carried out for noun and adjectives as well. The feature extractor divides these words in to parts of speech and saves all parts of speech in separate columns in database against the documents. Hence the documents can be compared on basis of their parts of speech features later which allows us to understand meaning of sentences as well.

C. *Synonyms Extraction*

NHunspell is a word stemming and thesaurus library based on the Open Office. This tool allows obtaining synonyms from WordNet. The NHunspell library takes keywords as an array for input and plots onto WordNet graph. The WordNet graph requires words to be in their root form otherwise they are not detected or recognized as nodes. So, this stemming process is done in the start. Now these stemmed keywords are plotted as WordNet nodes and their synonyms are retrieved. The graph provides synonyms against each keyword from Wordnet. These synonyms are extracted, and the software develops a separate dictionary in database and synonyms against each keyword in separate columns.

Finding synonyms against each keyword and storing into keyword dictionary by using following command.

```
syn = S.GetSynonymsBySystem(keyword);
string query = string.Format("insert into
dbo.KeywordsDictionary(Keywords,Synonyms)values('{0}','{1}')
", keyword, syn[]);
```

D. *Measuring Similarity*

For comparing the documents and calculating the similarity between them, we used SQL database of “Semantics DB” developed by Microsoft. SQL database of semantic search builds upon the existing full-text search feature in SQL Server, but enables new scenarios that extend beyond keyword searches. While full-text search lets you query the words in a document, semantic search lets you query the meaning of the document.

Semantics db takes input of synonyms and keywords from two documents and assigns weights to each word. The documents are then compared, and similar words are extracted.

The weights of these similar words are added, and a score of similarity is obtained. This score is multiplied by 100 and then saved in database system as similarity percentage.

Following command is used in query for passing document content to SQL semantics db,

```
documentsSemanticCal(Doc_Name, Doc_Content)
```

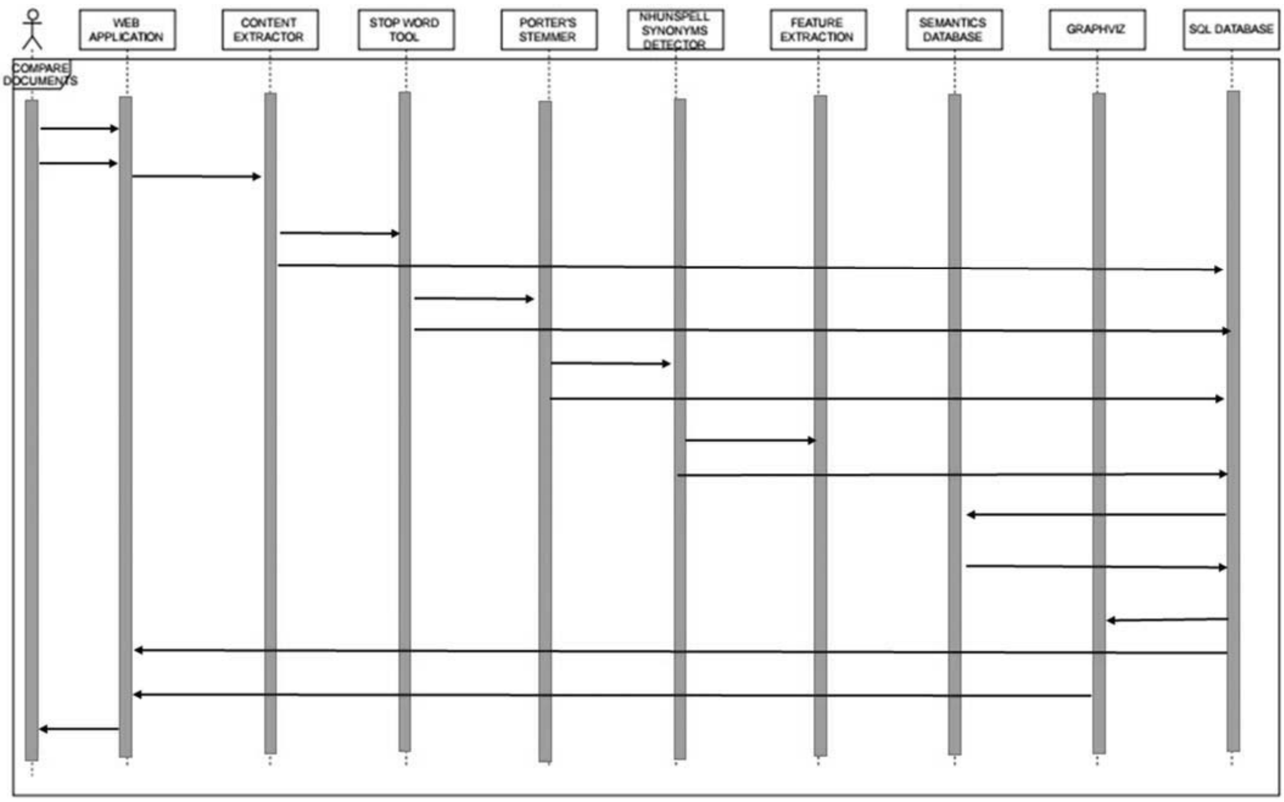

Fig. 2 Workflow diagram

E. Visualization

Graphviz library is used for visual represenatation of data. It takes keywords and synonyms from both documents as input nodes. Both documents' words are represented with nodes of different colors and similar words between them are assigned a separate color.

The graphical representation of how documents are related semantically is shown by integrating graphViz library in visual studio C#,

```
string graphVizString = @"digraph g {"+ abcc + ";}";
Bitmap ab = new
Bitmap(Graphviz.RenderImage(graphVizString, "jpg"));
```

III. Conclusion

The similarity between two documents can be calculated by various methodologies. We took the working of WordNet into account and by using semantic database, generated by Microsoft, we constructed a new software. This software is a user-friendly interface which takes two documents as input, calculates the relatedness between texts and then provides a visual description of how these documents are relatable. It is hoped that in future research, the shortcomings of this work are tackled and a ready to use high quality software is designed for direct use.

IV. Related Work

Taking the topic of semantic similarity into view, we can understand that the requirement is to extract extent of commonalities between two documents.

As per research upon semantic similarity and relatedness, we were not clear about the concept being similar or interrelated. It is hard to distinguish between relatedness and similarity. Two words are semantically related if they tend to be used near each other, and they're semantically similar if they tend to be used to replace each other; semantic similarity is more specific as compared to relatedness. Researcher [1] believed that semantic similarity and semantic relatedness are intimately related. Similarly, another research [2] specifies the difference between semantic similairty and relatedness in terms of their calculation method. Semantic relatedness can be calculated with usage of vector space model, and similarity metric or conventional metrics (e.g. cosine metric). It is measuring tendency of usage words together in sentences (documents, paragraphs). We can also explain this minute difference with a non-technical example. There is an example for clearing the concept of similarity and relatedness: 'ice cream' and 'spoon' both are related to each other because we eat ice cream with spoon but they are not similar because one cannot be used in place of other word.

Existing approaches for computing semantic similarity can be classified into two basic categories: corpus-based approaches and ontology-based approaches (knowledge-based approach). Both approaches enable to evaluate semantic interpretation of terms.

Researcher [3] used semantic distance (the Cilibrasi and

Vit´anyi's one) to compute the relatedness between two plain words. On the other hand, they have worked on ontology based semantic relatedness measurement, to measure word senses.

While making use of ontology-based similarity measure, the similarity between set of word pairs is calculated [4] and semantic similarity is calculated between concepts in an ontology by using WordNet. WordNet similarity measure is based in the lexical database, relatedness between concepts is measured by calculating cosine between a pair of vectors contain words in [5].

Researcher [6] has applied corpus-based approach, on WordNet for automatically acquiring domain name by comparing domain name Reuter's corpus with domain annotations already present in WordNet Domains.

We also considered which of the tools or methods were getting more focused by researchers for futuristic developments. A hybrid WordNet-based approach to measure concept semantic similarity has been proposed, which supports the information content of concepts to measure the cost of the shortest path distance between ideas (concepts) [7].

Furthermore, researchers in [8] made use of WordNet (A lexical database) for finding the semantic relatedness of geographical terms. Terminologies are mapped on the WordNet tree, a tree of words and results in semantic relatedness of any two words in the thesaurus.

Existing researches are based on semantic similarity between sentences, texts or concepts. However, document-level semantic similarity is not yet introduced. Researcher proposed a method that focuses on document-level semantic similarity which will be useful at academic institutes based on articles using top event and ontology [9].

The proposed model in [10] measures ontology based semantic similarity by using semantic distance between concepts, concept level and overlapping degree between sets of hypernyms and hyponyms whereas, existing ontology based measures focuses on semantic distance only [7], [9].

Acknowledgment

We had to take the help and guideline of some respected persons, who deserve our greatest gratitude. We would also like to thank our advisor (Dr. Faisal Fayyaz, Computer Scientist, Air University E-9 Islamabad, Pakistan) for guiding and consistently steering us in the right the direction.

We would also like to show our gratitude to the (Dr. Shahzad Hassan, Senior Assistant Professor, Bahria University E-8 Islamabad, Pakistan) for sharing their pearls of wisdom with us, and we thank to reviewers for their so-called insights.

Moreover, we would like to thank our parents who constantly supported and backed us in the difficult times.

References

[1] Slimani, T. (2013). Description and Evaluation of Semantic similarity Measures Approaches. 10.

[2] E. G., & S. M. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. 6.

[3] J. G., & E. M. (2008). Web-Based Measure of Semantic Relatedness. 15.

[4] Y. L., Bandar, Z. A., & D. M. (2003). An Approach for Measuring Semantic Similarity between Words Using Multile Information sources. 12.

[5] T. P., S. P., & J. M. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. 2.

[6] B. M., C. S., G. P., & A. G. (2002). Comparing Ontology-Based and Corpus-Based Annotations in WordNet. 6.

[7] Y. C., Q. Z., W. L., & X. C. (2017). A hybrid approach for measuring semantic similarity A hybrid approach for measuring semantic similarity. 25.

[8] Zugang Chen, Jia Song & Yaping Yang (2018). An Approach to Measuring Semantic Relatedness of Geographic Terminologies Using a Thesaurus and Lexical Database Sources. 22.

[9] Liu, Ming & Lang, Bo & Gu, Zepeng. (2017). Calculating Semantic Similarity between Academic Articles using Topic Event and Ontology.

[10] Y. Yang and Y. Ping, "An Ontology-Based Semantic Similarity Computation Model," 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, 2018, pp. 561-564. doi: 10.1109/BigComp.2018.00096

[11] Miller George, WordNet: a lexical database for English, Communications of the ACM, vol. 38, no. 11, pp. 39-41, 1995.

[12] J. Xu, Y. Tao, H. Lin, "Semantic word cloud generation based on word embeddings", 2016 IEEE Pacific Visualization Symposium (Pacific Vis), pp. 239-243, 2016.