IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A SUMMARY OF SEMANTIC SIMILARITY MEASURES BETWEEN WORDS

Pooja Tiwari

Research Student
Department of Computer Science
Jharkhand University of Technology, Ranchi, India

Abstract: Semantic similarity measures play a crucial role in various natural language processing tasks, aiding in tasks such as information retrieval, text classification, and semantic search. This paper provides a comprehensive review of semantic similarity measures, examining their methodologies, applications, and performance. We discuss different approaches to measuring semantic similarity, including knowledge-based, corpus-based, and hybrid methods, highlighting their strengths, limitations, and comparative evaluations. Additionally, we explore the challenges and future directions in semantic similarity research, aiming to provide insights for researchers and practitioners in the field of natural language processing.

***Index Terms* -** Semantic similarity, natural language processing, knowledge-based methods, corpus-based methods, hybrid methods.

I. INTRODUCTION

Measures of semantic similarity quantify how similar or related two texts are to one another in light of their underlying meaning. They have become indispensable tools in various natural language processing applications, including information retrieval, text mining, and ontology alignment. The accurate measurement of semantic similarity enables systems to understand and process language more effectively, leading to improvements in tasks such as document clustering, question answering, and recommendation systems.

Semantic similarity measures have mostly been developed to measure the degree of similarity between two words or two concepts utilizing pre-existing resources that store associations between words or concepts, such as the WordNet lexical database.[[1], [2]]

The challenges in achieving semantic similarity across different levels of representation, emphasizing the importance of operating at the sense level to overcome limitations in existing methods. The proposed unified approach offers advantages in enabling meaningful comparisons across various scales of text and identifying semantic similarities beyond lexical forms and ambiguity.

A sense-based representation is crucial for accurately detecting similarities in meaning between words, even when there is minimal lexical overlap. This approach is particularly important when different words are used to express the same idea.[3]

The Traditional approaches to computing semantic similarity between terms using manually compiled dictionaries like WordNet are limited because they do not cover all terms, such as abbreviations and brand names. The text suggests utilizing Web Search Engine (WSE) based approaches, such as Google and Bing, to leverage collective intelligence and solve problems related to semantic similarity. The goal of the article is to explore and estimate different methods, including leveraging past search trends from WSE, that can intelligently quantify the similarity between new terms that are not often covered in dictionaries.[4]

The OWL (Web Ontology Language) and Service ontology (OWL-S) is now the standard for developing semantic service descriptions. By thinking about the different service kinds, one can determine service compatibility. Let Service B, for instance, be a subclass of Service A. When a composition scenario calls for a type A service, Service B can be used in its place to finish the composition. Ontology concepts define functionality, and services can be matched by their ontological annotations through composition based on capability requirements. The current method of logical reasoning offers a formal model for composition that is automatic. However, this kind of logical approach is sometimes too restrictive in scenarios when fully automated composition is not needed (or even desirable). For instance, take into consideration a user-friendly service composition environment where intricate workflows can be created. Users would rather have a selection of "similar" services that could be built in this kind of semiautomatic composition system than one or two precise logical matches. In order to match OWL-S annotated services, we present a method in this work for assessing the similarity between OWL objects (classes and instances). A comparison of the semantic descriptions of two services is used to determine how comparable they are. Since each service description consists of a collection of RDF statements, we may define "similarity" between services by calculating the ratio of the description statements that are shared by the two services. We are calculating the ratio between the total number of descriptions and the number of common statements in both service descriptions. It makes sense that services are more comparable to one another if they share more information. A practical and lightweight method for utilizing the existing semantic metadata is to use the service similarity metric. It is uncommon for the computationally demanding method of logical reasoning to be employed in a large-scale heterogeneous distributed system (i.e., the Grid) to provide a suitable outcome within time constraints. We suggest doing an optimization phase using the similarity measure prior to any required logical deduction procedures. By generating a list of similar services ahead of time, this kind of optimization strategy is crucial for semantic service matching and helps to cut search time. Then, on demand, the required logical reasoning can be carried out. Semantically optimized service To locate the collection of composable services in a real-time service composition environment, search is necessary.[5]

II. Related Works

The Author in [5] paper introduces a numeric metric for calculating semantic similarity in OWL Lite ontology by determining shared descriptive information between objects. The focus is on comparing OWL-S annotated services using a weighted aggregate of Service Profile, Service Model, and Service Grounding, showcasing a method for matching semantic services based on similarity measure. Ongoing implementation aims to apply the information-based similarity algorithm to various application scenarios, particularly in the semantic matching of grid services, with a focus on effectiveness compared to inference-based methods and exploring the impact of different rule sets on the similarity metric's quality.

In conclusion, Author of this paper[6] discussed the importance of properly processing various information sources in defining similarity measures for words, highlighting the limitations of WordNet for new words. Different researchers' approaches were presented, with the Boll gala, Matsuo, and Ishizuka (BMI) approach showing the highest correlation of 0.87, indicating superior performance compared to other methods. Moving forward, incorporating more information resources or their combinations is suggested to enhance the measurement of semantic similarity between words.

The work in [3] achieves state-of-the-art performance in three experiments and proposes a unified approach for computing semantic similarity at several lexical levels. Future research will examine the effects of sense inventory-based networks and assess the technique on longer textual units.

In text processing and information retrieval, semantic similarity is essential. The semantic similarity metric is used to determine how similar words, phrases, and documents are to one another. This Paper[7] reviewed semantic similarity in its entirety, discussed its uses, and provided an explanation of how to measure it. Semantic similarity measures base their semantic similarity assessment on knowledge bases like WordNet, Wikipedia, and Ontology. The semantic similarity has been categorized in this evaluation according to measurements and strategies, and their benefits and drawbacks have also been covered.

We provide four distinct STS measures in this[8] survey. We categorize the String similarity measurements into two groups: term-based and character-based. By measuring the distance between two strings, sets, or vectors, these two metrics analyze similarities and differences in string sequences and character compositions. In total, we provided fourteen string similarity metrics that were divided into two categories.

Three types of topological approaches were also covered: hybrid, which combines node and edge bases, edge-based, and node-based techniques. The primary purpose of these topological research is to identify terminology and ontological notions that are similar and different from one another.
Additionally, it was found that node-based approaches are fully dependent on the information content value between two nodes, but distance-based approaches depend on the depth of the semantic network. Conversely, the hybrid approach uses the weight value between the parent and child nodes to determine how similar the two classes are. We described six distinct techniques in the statistical approach, including LSA, GLSA, ESA, PMI-IR, NGD, and HAL. By reducing the dimensionality of the vectors and generating a vector space model from the corpus, these statistical measures identify similarities and differences.
Every cell in the matrix denotes the frequency or weight value of a specific word in a text passage or paragraph.

The author in paper[4] describe use of semantic similarity measures is important in various applications, and the evaluation of novel techniques using knowledge from WSE has shown promising results. Future work should focus on avoiding cognitive bias and reaching a common agreement on data evaluation, as well as exploring the use of automatic construction of ontologies to improve accuracy in semantic similarity measures.

An important component of many applications within the field of artificial intelligence research is the evaluation of semantic similarity is discussed in[9] . Various approaches can be distinguished based on the theoretical underpinnings and the manner in which ontologies are examined to calculate similarity. In order to evaluate the similarity of concepts or phrases, this study offers a sophisticated analysis of the most widely used semantic similarity metrics. The purpose of this study is to provide some insights on the accuracy, typology, and essential characteristics of the measurements that are discussed under each category. Furthermore, an effective comparative analysis of all these metrics within an actual context is provided, utilizing the two often used reference points. The advantages deduced from those studies would assist practitioners and researchers in choosing the measure that most closely matches the needs of a practical application.

Author in [10]shows survey of publications in three categories, including knowledge-based, corpus-based, and string-based methods, and a survey of semantic similarity methods. A thorough examination and analysis are carried out on 25 papers, utilizing diverse techniques and methodologies to gauge semantic similarity. The analysis reveals that approaches based on knowledge and corpora are frequently employed to measure semantic similarity and produce encouraging outcomes.

There are numerous uses for the semantic similarity score between a given pair of texts. The similarity between the two texts' embeddings is measured to provide this semantic similarity score. In this research[11], Many custom-made bag-of-words based systems that compute the semantic similarity are constructed using word embedding techniques such as Word2vec and GloVe. To obtain better embeddings from the text, bag-of-words based techniques also make use of concepts like TF-IDF, Word Mover's distance, and Smooth Inverse Frequency. Infer sent and other pre-trained encoder models are also employed to create the embeddings.

Additionally, a new model is developed to create embeddings from input text. It is based on a foundation of a Keras ensemble of Universal Sentence Encoders and evolves over several layers into a deep learning model trained on the SICK-train and SICK-dev datasets to calculate the semantic similarity score between two texts. The different methods are assessed using the cosine similarity function between the generated embeddings for sentence pairs using the SICK-test dataset.. Pearson, Spearman, and Kendall's Tau correlation measures are used to compare the predicted similarity values with the ground truth values. According to experimental data, the innovative model performs better than all other models, indicating that it may be effectively utilized to extract semantic information from the input text. Semantic similarity has many possible applications. The best performing method (the novel model) is used to build a proof-of-concept application that finds the semantic similarity between a set of documents. Additionally, the suggested work can be applied to a number of disciplines, including general academics, literature, and medical. Incorporating methods to assess the degree of resemblance between collections of photos embedded in other publications in the future would also be intriguing.

In conclusion,[12] this study focused on sentence semantic similarity using WordNet, comparing different measures and exploring theoretical properties. The research highlighted the importance of PoS tagging in determining sentence similarity and proposed a new method using PoS WordNet-based conversion, which outperformed existing techniques. Future work will involve incorporating named entities and semantic role labelling to enhance sentence similarity assessment, as well as considering the use of Concept Net in place of WordNet for lexical resources.

.

III. RESEARCH METHODOLOGY TECHNIQUES FOR MEASURING SEMANTIC SIMILARITY

Measuring Semantic Similarity Based on information sources, methods are employed to determine the semantic similarity of words. Web search engines are one type of information source, along with ontologies like WordNet, biomedical dictionaries, and Brown Corpus. Some techniques based on both sources are listed below.

3.1 Traditional Ontology based methods

Semantic similarity assessment techniques based on ontologies are those that employ ontologies as their main source of information. They can be loosely divided into the following three groups:

3.1.1. Distance based method

A more straightforward and natural method of assessing the semantic similarity of terms using taxonomy is the distance-based approach. It calculates the separation (such as edge length) between nodes that represent the ideas under comparison. The geometric distance between the nodes representing the concepts can be used to conveniently estimate the conceptual distance in the given multidimensional concept space. Rada et al.[13] stated that in order for a taxonomy to be hierarchical, the distance must meet certain metric qualities, including positive, zero, symmetric, and triangular inequality.
As a result, in an IS-A semantic network, the shortest path—that is, the least number of edges separating two basic idea nodes, A and B—is the easiest way to calculate their distance from one another.
When Rada et al. [13]used the distance approach in the medical field, they discovered that the distance function accurately represented how people would evaluate conceptual distance. Richardson and Smeaton, however, expressed worry that the measure's accuracy was lower than anticipated when it was applied to a very broad domain, such as WordNet taxonomy. They discovered that unexpected conceptual distance outcomes are caused by irregular density of linkages between ideas. Additionally, because of the general structure of the taxonomy, the value of the depth scaling factor does not change the overall measure well, without having numerous major side effects elsewhere.
Furthermore, we believe that the subjectively pre-defined network hierarchy plays a major role in determining the distance measure. A few local network layer constructs might not be appropriate for direct distance manipulation because the primary goal of the WordNet's design was not to compute semantic similarity.[6]

3.1.2. Information content based method

Resnik in [14] noted that the information content based method refers to the node-based methodology used to assess conceptual similarity. The amount of information that two words have in common determines how similar they are. This is done in a multidimensional space where each node represents a distinct concept with a specific amount of information and each edge always represents a direct relationship between two concepts. This shared information "carrier" can be recognized in a hierarchical concept space as a particular idea node that encompasses both of the two words. To put it simply, this super-class should be the class that comes before all other classes in the hierarchy that includes both classes. The information content value of this particular super-ordinate class is the definition of the semantic similarity value. Next, a class's information content value is determined by calculating the likelihood that a certain class will appear in a sizable corpus of texts.
Less information on the intricate taxonomy structure is needed for the information content method. It is not affected by the issue of different link kinds. Nevertheless, as it disregards structure-related information, it remains reliant on the taxonomy's basic framework. Typically, it produces a rough outcome when comparing words. This means that, provided that two concepts have the same "smallest common denominator," it does not distinguish between the similarity values of any pair of concepts inside a subhierarchy.

Lin [15]uses an information-theoretic formula to determine semantic similarity. Lin's change involved supposing the ideas' independence and normalising them based on the combination of their information content. Lin's[15] similarity measure employs an information-content methodology predicated on three suppositions. First of all, this notion will share more in common with other concepts the more similar they are. Second, there is a decreasing degree of similarity between two concepts the less they share. Thirdly, when two conceptions are the same, their resemblance is at its highest.

3.1.3. Distance and Information Content based method

A method for calculating the semantic similarity of words and concepts was proposed by Jiang and Conrath[16] .In order to better quantify the semantic distance between nodes in the semantic space created by the taxonomy using computational evidence obtained from a distributional analysis of corpus data, it integrates corpus statistical information with lexical taxonomy structure. To put it briefly, this methodology combines the node-based information content measurement with the edge-based edge counting scheme to create a combined approach that is improved.

Semantic distance is provided by the Jiang-Conrath measure as opposed to similarity or relatedness. By taking the multiplicative inverse of this distance measure, one can transform it to a similarity measure.

3.2 Web Search Engine (WSE) Based Approaches

Semantic similarity analysis has become a vital component of several domains, including natural language processing and information retrieval. We are attempting to quantify the semantic similarity between two provided words, w1 and w2, in order to solve the problem. Measuring the innate similarities between two or more concepts is known as similarity. Beyond synonymy, semantic similarity is a notion that is frequently defined in the literature as semantic relatedness [15].

Bollegala et al. noted that there was some semantic resemblance between meronyms (book, page) and hyponyms (rose, flower), in addition to synonyms (noon, noon).

Explicit semantic analysis (ESA) provided a fresh approach for information retrieval studies and associated research. This method measures the semantic relatedness between two words in concepts space rather than a terms space, so the relationship is assessed in both the text's lexical form and the words' meanings.

To calculate the semantic similarity between words, we use the ideas space rather than the terms space; that is, we compare the meanings of terms rather than the lexicography associated with them. For instance, from a lexicographical perspective, the terms "hat" and "rat" are quite similar, but they do not convey the same meaning. Since a similarity score of 0 indicates total inequality and a score of 1 indicates complete equality between the concepts being compared, our focus is solely on the real-world concept that they reflect.

A great deal of research has been done on evaluating semantic similarity using Web material over the years.

Semantic similarity measurements that employ WSE-based techniques fall into one of the following categories:

- Snippet based methods
- Page count based co-occurrence measure methods
- Frequent pattern finding based methods
- Trend Analysis based methods

3.2.1. Snippet based methods

These strategies involve gathering the text excerpts produced by search engines such as Google immediately produced the result after conducting a search for these phrases. These text excerpts can be used to compare different algorithms that calculate the semantic similarity between two phrases based on the text excerpts that are related to each other.

Using the excerpts that a WSE gave for two searches, Sahami et al.[17] presented a semantic similarity metric. Snippets are gathered from a WSE for every query, and they are represented as a weighted word vector with the terms "Term Frequency" and "Inverse Document Frequency" reversed. This strategy captured more of the semantic context-based similarity measures in the collected snippet than in taxonomy-based similarity measurements. Consequently, there is a difference in the amount of semantic similarity between high TF and IDF. This method's primary flaw is that it can only process a query's top-ranking results quickly. A method of double-checking with text fragments returned by the WSE was proposed by H. Chen et al.[18] .

If one can be located using a web search engine from the other, then the two objects are said to be connected. Take the Co-Occurrence, for instance. The main disadvantage of this approach is that, despite their relationship, we cannot guarantee that a word will appear in the snippets for the other event.

3.2.2. Page count based co-occurrence measure methods

It involves calculating the likelihood that the terms will occur together on the Web based on the number of pages. When the two words w1 and w2 are supplied as input, WSE returns page counts for them. Hits on the Web indicate the likelihood of a word co-occurring. These formulas are assessments of the likelihood that the phrases w1 and w2 will occur together. The number of hits returned when a certain WSE is provided with this search term divided by the total number of web pages that might be retrieved indicates the likelihood of a given term. The combined probability, or p (w1, w2), is the number of hits a WSE produces that contain both search keywords w1 and w1, divided by the total number of web pages returned[19].
In this sector, Normalised Google Distance (NGD)[19] is regarded as one of the best studies. The NGD is a measure of semantic similarity for a particular group of phrases that is derived from the Google search engine (GSE).

3.2.3. Frequent pattern finding based methods

The methods used by this group fall under the category of machine learning and involve searching websites indexed by a certain WSE for patterns of similarity. Bollegala et al.[20] introduced a well-known method that involves searching for regular expressions like "w1 is w1 w2," "w1 is also known as w2," "w1 is an example of w2," and so on. This is due to the fact that this type of statement highlights how similar the two (set of) terms are semantically.
The resemblance between the two phrases is supported by a high frequency of these kinds of patterns, but we must first conduct some preliminary research to determine what constitutes "a high number" in relation to the issue we hope to solve. This can be achieved, for instance, by looking at how many results a specific WSE, like Google, returns when you search for ideal synonyms. Additionally, since the similarity between w1 and w2 is equivalent to the similarity between w2 and w1, it is imperative to consider that these expressions should be examined in two different ways.

3.2.4 Trend Analysis based methods

Time series are collections of well-defined data items that are measured repeatedly, and techniques based on trends analysis are used to find an underlying pattern of behaviour in the series. WSE stores the queries in this way so that it can use the data in the future. Over time, many methods have been proposed to compute semantic similarity and evaluate the correlation between search patterns. Using Pearson's correlation coefficient, which is closely connected to the Euclidean distance over normalize vector space, Jorge Martínez Gil in [21]has presented a method. Rather of giving numerical numbers, this measure gives the mean as a time series.
As a result, there is a very high likelihood of semantic resemblance between the related concepts, which may have nearly identical shapes in their related series.

IV. Conclusion

These days, the semantic similarity metric is crucial to a lot of applications. First, we have studied a number of ontologies that are employed for semantic similarity in this survey. This is the conventional approach of measuring semantic similarity. Second, we have discussed and assessed a number of innovative and promising methods for assessing the semantic similarity between words that make use of WSE information. Utilizing a benchmark dataset of phrases that are not frequently found in dictionaries, taxonomies, or thesaurus, all of the methodologies reviewed have been assessed. Consequently, we have verified through experimentation that certain WSE-based techniques much surpass current approaches when assessing this type of dataset.
In further research, the richness of the fact that individuals assess our word pairings in a variety of ways based on their cultural background will be a determining factor in how successful semantic similarity is.

V. Future Scope

Thus, going forward, we want to steer clear of cognitive bias when it comes to phrases that, while synonymous to one person, may not be so to someone from a different culture (and vice versa). We must come to a consensus regarding the data that is utilised to compare various strategies. It's possible that using automatic ontology creation (KBs) will lead to future developments that improve the semantic similarity measure's accuracy.

The latter relates to the requirement for information sharing among agents during agent communication. The main objective is to determine the best solutions for this problem and implement them in real information systems where it would be necessary for the automatic calculation of semantic similarity between phrases.

References

[1] George . (2000). WordNet: An Electronic Lexical Database. Vol. 76. pp. 706-706. , https://doi.org/10.2307/417141

[2] George . (2000). WordNet: An Electronic Lexical Database. Vol. 76. pp. 706-706. , https://doi.org/10.2307/417141

[3] Mohammad. (2013). Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. Vol. 76.pp. 1341-1351. , https://doi.org/10.2307/417141

[4] Ali, Ashraf & Alfayez, Fayez & Alquhayz, Hani. (2018). SEMANTIC SIMILARITY MEASURES BETWEEN WORDS: A BRIEF SURVEY. 30. 907-914.

[5] Jeffrey. (2005). A Semantic Similarity Measure for Semantic Web Services. Hau, Jeffrey & Lee, William & Darlington, John. (2005). A Semantic Similarity Measure for Semantic Web Services.

[6]Ankush. (2012). Measurement of Semantic Similarity Between Words: A Survey. Vol. 2. pp. 51-60. , https://doi. org/10.5121/ijcseit.2012.2605

[7] D., Akila. (2018). Semantic Similarity- A Review of Approaches and Metrics. International Journal of Applied Engineering Research. 9.

[8] Majumder, G., Pakray, P., Gelbukh, A., & Pinto, D. (2016, December 26). Semantic Textual Similarity Methods, Tools, and Applications: A Survey. Computación Y Sistemas, 20(4). https://doi.org/10.13053/cys-20-4-2506

[9] Slimani, T. (2013, October 18). Description and Evaluation of Semantic Similarity Measures Approaches. International Journal of Computer Applications, 80(10), 25–33. https://doi.org/10.5120/13897-1851

[10] P, Sunilkumar & Shaji, Athira. (2019). A Survey on Semantic Similarity. 1-8. 10.1109/ICAC347590.2019.9036843.

[11] Agarwala, Saurabh & Anagawadi, Aniketh & Guddeti, Ram. (2021). Detecting Semantic Similarity Of Documents Using Natural Language Processing. Procedia Computer Science. 189. 128-135. 10.1016/j.procs.2021.05.076.

[12] Oussalah, M., & Mohamed, M. (2021, August 21). Knowledge-based sentence semantic similarity: algebraical properties. Progress in Artificial Intelligence, 11(1), 43–63. https://doi.org/10.1007/s13748-021-00248-0

[13] Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1), 17–30. https://doi.org/10.1109/21.24528

[14] Resnik . (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. , https://doi.org/10.48550/arxiv.cmp-lg/9511007

[15] Lina . (1998). An Information-Theoretic Definition of Similarity. pp. 296-304. ,

[16] Jinag. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. , https://doi.org/10.48550/arxiv.cmp-lg/9709008

[17] Sahami. (2006). A web-based kernel function for measuring the similarity of short text snippets. , https://doi.org/10.1145/1135777.1135834

[18] Chen . (2006). Novel association measures using web search with double checking. , https://doi.org/10.3115/1220175.1220302

[19] Cilibrasi . (2007). The Google Similarity Distance. Vol. 19. pp. 370-383. , https://doi.org/10.1109/tkde.2007.48

[**20**] Bollegala . (2007). Measuring semantic similarity between words using web search engines. , https://doi.org/10.1145/1242572.1242675

[21] Martinez-Gil . (2012). An overview of textual semantic similarity measures based on web intelligence. Vol. 42. pp. 935- 943. , https://doi.org/10.1007/s10462-012-9349-8