An Approaches & comprehensive survey for Measuring Semantic Relatedness with Knowledge Resources

[1]Sunita, [2]Dr. Vijay Rana

[1]Research Scholar in CS Department, [2]HOD & Prof. in CSA Dept
Arni University Kathgarh Indore(HP), SBBS University Punjab(PB)

***Abstract*: Semantic relatedness is a type of estimation that quantitatively recognizes the connection between two words or ideas in light of the comparability or closeness of their importance. In the current years, there have been critical endeavors to register SR between sets of words or ideas by misusing different information assets, for example, semantically organized (e.g. WordNet) and cooperatively created learning bases (e.g. Wikipedia), among others. The current methodologies depend on various strategies for using these learning assets, for example, techniques that rely upon the way between two words, or a vector portrayal of the word portrayals. The motivation behind this paper is a comprehensive survey Measuring Semantic Relatedness between Web Search Content. Furthermore, we give rules to analysts and specialists on the best way to choose the most important SR strategy for their motivation. At last, in light of the similar investigation of the looked into relatedness measures, we distinguish existing difficulties and conceivably important future research bearings in this space.**

***Keywords*: semantic relatedness, semantic similarity, information based measurement, information content.**

Introduction

Measures of relatedness or likeness are utilized as a part of an assortment of uses, for example, data recovery, programmed ordering, word sense disambiguation, programmed content amendment. Semantic closeness and semantic relatedness are in some cases utilized exchangeable in the writing. These terms in any case, are not indistinguishable. Semantic relatedness demonstrates degree to which words are related by means of any sort, (for example, synonymy, meronymy, hyponymy, hypernymy, practical, acquainted and different composes) of semantic connections. Semantic similitude is an extraordinary instance of relatedness and thinks about just hyponymy/hypernymy relations. The relatedness measures may utilize a mix of the connections existing between words relying upon the unique situation or their significance[18]. To outline contrast amongst comparability and relatedness, Resnik [1] gives the broadly utilized case of auto and fuel. These terms are not fundamentally the same as; they have just couple of highlights in like manner. In any case, they are all the more firmly related in a useful setting; specifically that autos utilize gas. Various analysts utilize separate measure as inverse of comparability.

Human can regularly easily choose about the closeness or relatedness of two words. This can be clarified, to some degree, by the experience that people have in utilizing and experiencing related words in comparative settings. For example, as individuals, we know rain and umbrella are profoundly related, while there is a bit, assuming any, association amongst rain and course reading. While this is unimportant for people, it is frequently not as easy to decipher this judgment procedure for machines without the cautious detailing of foundation and logical information encompassing each word and its connections. Formally, semantic relatedness (SR) is characterized as a type of semantic or practical relationship between two words as opposed to simply lexical relations, for example, synonymy and hyponymy (Budan and Graeme, 2006). The target of SR techniques is to firmly model such affiliations. SR is widely used in many practical applications, particularly in natural language processing (NLP) including semantic information retrieval, keyword extraction and document summarization, where it is 1 While acknowledging the differences, we use the terms 'words, concepts, terms and entities', interchangeably in this paper[19], used to quantify the relations between words or between words and documents (Leong & Mihalcea, 2011).

2. Semantic relatedness methods

Our research objective is to develop a framework that allows us to compare some of the well-known methods in the SR literature. We adopted an iterative approach towards the design of this framework.We initially base our work on the three main dimensions that have already been highlighted in the literature (Agirre et al., 2009;Chen et al., 2009), namely knowledge resources, computational methods and evaluation approaches[20]. We then identify several importantwork in the literature that would be considered seminal or novel work in the domain of SR. Our criteria for selecting these methods are as follows:

Selecting methods with a substantial impact on the literature: Our objective has been to select and review methods that have had a notable impact on the research community. For this purpose, one of the criteria for choosing a study has been its citation count obtained through Google Scholar. We postulate that the higher the citation count for a publication, the better the proposed method has been received and recognized by the community.

Selecting methods with original proposals: Our goal has been to include work that was the first to propose an idea with regards to using a knowledge resource or a computation method. The selection included studies that were original work in proposing the idea and not adoptions of earlier ideas. To decide on originality of two similar pieces of work, the work published earlier and cited by other work in an earlier chronological order was chosen as the original one.

Research Gap

Authors	Objective	Technique	Characteristics	Limitation
Cilibrasi	Find the semantic distance between two contexts	Normalized google distance	Convert the query based research into knowledge based research	Redundancy Issue
Zhong n	Internet of things accelerate	Semantics model	An environment to provide active ,transparent, safe and reliable services	Ambiguity problem
Andrea moro	Disambiguation the words sense	Tokenization	Remove the ambiguity of word	Redundancy issue
Mohammed maree	Find semantic correspondence between ontologies	Ontology matching	Conquer the semantic heterogeneity problem	Domain specific knowledge
Ensan and Bagheri	Relatedness of a query to a given document is Calculated.	Concept graph using DBpedia.	Represents documents and queries through a graph of concepts.	Focus only on query of document not on main concept
Ganggaozhu	Method for measuring the semantic similarity between concepts in knowledge graphs.	W-path	Compute semantic similarity	Does not compute relatedness
Phillip Resnik	shared information content value of two words based on subsumption relations	IS-A Taxonomy	Measure performs better than the traditional edge-counting approach.	Only work Supervised Learning Technique
Gracia	proposed web based semantic relatedness technique	Normalized Google Distance (NGD)	measure to compute the relatedness degree of co-occurrence of words on web pages	co-occurrence of words on web pages
Miller	a lexicalized concept	WordNet3.0	Semantic relations link the synonym sets	WordNet is an online lexical database designed for use under program control.

Literature Review

Literature survey plays an imperative role in our research work. It is the documentation of a comprehensive review of particular theme, which holds the information of past and present development of the topic. Thus it motivates to develop innovative techniques and models. This work describes the work of eminent researchers and highlights the challenges, which still require to be addressed.

Resnik (1995) hypothesizes that SR between two words is a measure of the amount of information they share. For this purpose, and in order to identify shared information, the method proposed in Resnik (1995) identifies the lowest common subsumer of the two words within an IS-A hierarchy. The information content value of the subsumer is regarded as an indicator of SR.

Jiang and Conrath (1997) employ the information content value of words as well as the information content value of the two words' lowest common subsumer in a lexical taxonomy structure to compute SR. The information content value of two words'

lowest common subsumer describes the amount of information these two words share, whereas the information content value of a word indicates how informative that specific word is. Here, SR is defined based on the information content of the lowest common subsumer in the context of the information content of each individual word.

Lesk (1986) structures his work on the short pieces of text (glosses) defining each word in WordNet. Specifically, SR is computed by counting the number of word overlaps in the glosses of the two words, where higher overlap means higher relatedness between two words.

Cilibrasi and Vitányi (2007) have proposed a method that relies on the information retrieved from a Web search engine. The motivation behind their work is that similar words when used as search queries will result in similar Web page results. Therefore, the count of the number of shared Web pages returned by a Web search engine for three different search queries, namely w1, w2, w1 and w2, is used to formalize the normalized Google distance (NGD). SR is defined as the inverse of NGD.

WikiRelate! (Strube & Ponzetto, 2006) takes advantage of Wikipedia articles and category tree to compute SR. In their work, the authors apply to Wikipedia the measures that were originally designed for WordNet. Articles are retrieved from Wikipedia by querying word pairs. Wikipedia's disambiguation pages obtained for each word are used for disambiguation of the words. The categories related to the retrieved articles are used to compute SR by for instance, considering the length of the shortest path or the length of the path that maximizes information content.

Sahami and Heilman (2006) have introduced a new approach for computing SR aimed at overcoming the poor performance of traditional document similarity methods when applied on short text snippets (Sahami & Heilman, 2006).

Their method, similar to the work in Cilibrasi and Vitanyi (2007), benefit from Web search results. In particular, they leverage Web search results for enhancing short snippets. Top ranked words, based on the TF–IDF measure from the search results, are used to build a vector for each input word. The vector is then used to compute the degree of SR between the two words.

Patwardhan and Pedersen (2006) used the co-occurrence information as well as the definitions of words in WordNet to build gloss vectors corresponding to each word. The gloss vector is created in two steps:

1. The first-order vector consisting of co-occurrences between the target word and other words among all the glosses in WordNet is formed.
2. additionally word co-occurrence information are calculated by concatenating the glosses of words that are related to each other within WordNet. Cosine similarity is applied to the gloss vectors to measure the relatedness between two words.

Their method, similar to the work in Cilibrasi and Vitanyi (2007), benefit from Web search results. In particular, they leverage Web search results for enhancing short snippets. Top ranked words, based on the TF–IDF measure from the search results, are used to build a vector for each input word. The vector is then used to compute the degree of SR between the two words.

Hughes and Ramage (2007) present an application of Markov chain theory to measure SR based on a graph extracted from WordNet. The graph is constructed such that the nodes are entries in WordNet and the edges are relational links between words. The authors adopted three types of nodes including Synset nodes, TokenPOS nodes and Token nodes, whereas the relationship types are hypernym/hyponym, instance/instance of, antonym, entails/entailed by, adjective satellite and causes/caused by. SR is calculated by assuming a particle that starts from a specific word, and then roams through the constructed graph.

The particle tends to explore the neighborhood related to the target word, hence resulting in a stationary distribution. SR is the similarity between two stationary distributions obtained for the two words.

Resources of Relatedness

In the previous section discussed several approaches of Semantic Relatedness. In this section we are discussed knowledge resources aspects of semantic relatedness, there is categorized by below

- **Knowledge resources**

In the context of SR techniques, the term knowledge resource refers to the type and source of information that is used for determining the degree of relatedness between two words.

a) Linguistically constructed resources (Relations, Synsets in WordNet and GermaNet).

WordNet

WordNet is a large lexical database for the English language. It consists of information that describes English words and expresses various meanings that a word can have in different contexts. Relations and synsets are two of the main constituents of WordNet where relations express information such as hypernymy,antonymy and hyponymy, and synsets represent groups of synonymous words. Additionally, members of each synset are often further described using a short piece of textual descriptor called the gloss. Various researchers have already benefited from Wordnet for computing the degree of SR between two words. These works have exploited both WordNet's relations and its glosses.

b) Collaboratively constructed resources (Articles, Article links, Categories, Disambiguation pages in Wikipedia, Information and Relations in English and German Wiktionary).

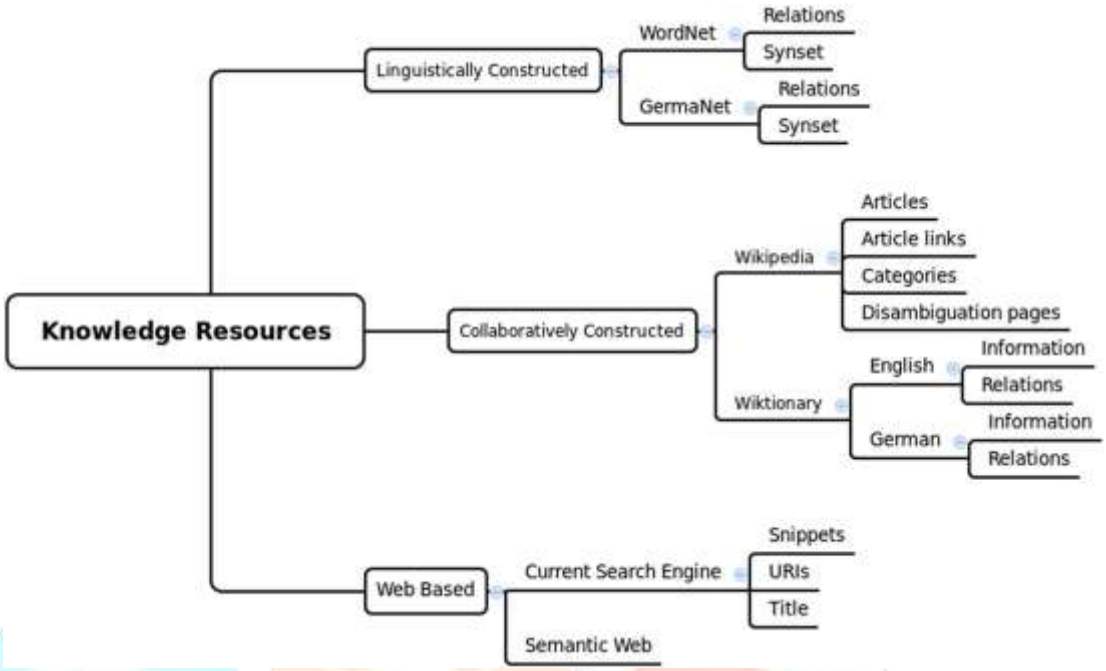
Wiktionary

Wiktionary is a multilingual, Web-based, freely available dictionary, thesaurus and phrasebook (Zesch et al., 2008) designed as a lexical companion to Wikipedia. Wiktionary shares many commonalities with Wordnet as they both include words, lexical relations between words and short pieces of text describing the words (glosses). Given the fact that Wiktionary consists of a large number of words, a high dimensional concept vector can be constructed based on its constituent words

Wikipedia

The information collected in Wikipedia is represented through the so-called articles, which are focused on and dedicated to the description of a specific topic. The content of each article is gathered and edited collaboratively and is often strictly moderated by

community volunteers. Besides articles, Wikipedia provides hyperlinks between articles, categories and disambiguation pages. Various researchers have already benefited from the textual content of Wikipedia articles, the hyperlink graph structure as well as categories and disambiguation pages to develop SR measures.



c) Web-based resources including Web search engine results.Web-based resources (Web Search Engines such as Google, Yahoo, Bing and the Semantic Web,i.e. the Linked Open Data cloud). In the Web-based knowledge resource category, two main information sources have been used, namely Web search engines and semantic Web resources.

Web search engines

Given the size of the Web and the role of search engines in content retrieval, there have been extensive research that have looked at how the results of search engines can be taken as an indication for SR. For a given search query, search engines often return useful information such as result snippets, Web page URIs, user-specified metadata and descriptive page titles[18]. The information content value of the outputs of search engines have been considered as possible indicators of relatedness. Web search engine snippets are short pieces of text for each result returned by search engine that contain a set of terms that describe the retrieved page. Some authors have benefited from snippets to measure SR[19]. For instance, Spanakis et al. (2009) have proposed a hybrid Web-based measure for computing SR between words by automatically extracting lexico-syntactic patterns from snippets based on the idea that similar words should have similar usage patterns. Similarly, Bollegala et al. (2007) have developed a SR method that relies on search snippets, and considers both word counts and lexicalsyntactic patterns when comparing the results of three queries w1, w2 and (w1 and w2). Sahami and Heilman (2006) collect snippets of the top ranked pages for a query and represent each query through an TF–IDF term vector of the collection of snippets. SR of two words is then computed based on the similarity of their query term vectors. Furthermore, Chen et al. (2006) have proposed a double-checking model to analyze snippets returned by a Web search engine, where the double-checking model is formed by a forward process which counts the total occurrences of w2 in the top N snippets of query w1 and a backward process which counts the total occurrences of w1 in the top N snippets of query w2. Duan and Zeng (2012) count the occurrences of each word and also the co-occurrence of the two words within the returned snippets and compute SR based on the obtained count frequencies.There have been other works based on Web search engine results that do not necessarily rely on snippets only, but also consider the content of the retrieved Web pages. The main reason for this is the short length of snippets that could impact the accuracy of the SR measures. For example, Sahami and Heilman (2006), who initially considered snippets as their knowledge resource, have enhanced snippets by adding the top-k words with the highest TF–IDF value from each of the returned document to the vector.

Semantic Web

recent knowledge resources are provided by the Semantic Web community in the form of ontologies and the Linked Open Data. These resources are based primarily on the RDF model, built of triples in the form of<subject, predicate, object>. A triple explicitly defines a relationship between a subject and an object through a meaningful relationship, known as a predicate. As introduced earlier, REWOrD (Pirró, 2012) is one of the earlier works that exploit the concept of Linked Open Data, especially the DBpedia knowledge base, to compute SR. In REWOrD approach, the correspondence between words and DBpedia's semantic concepts are first found. The retrieved DBpedia concepts are then used to construct a vector for each word. Vector similarity is used as the measure of SR between two words. Gracia and Mena (2008) have calculated SR between two concepts within a Semantic Web ontology by finding and comparing the similarity of their ontological contexts. An ontological context for a concept is defined as a collection of highly related concepts within the ontology that can support unambiguous definition of the given concept. For instance, the ontological context of a concept can include its hypernyms and synonyms. Karanastasi and

Christodoulakis (2007) have introduced OntoNL SR measure that depends on semantic relations defined by the Web Ontology Language. In this model, the authors compute SR by integrating three aspects: the number of common properties and inverse of properties that the two concepts share, the path distances of two concepts' common subsumer and the count of the common nouns and synonyms from the concepts' descriptions in the ontology

Association with human opinion

One of the fundamental procedures for assessing SR techniques has been to contrast their results and a best quality level informational index, for example, those presented prior. Scientists have either contrasted the outright anticipated relatedness esteem and the relatedness estimation of the best quality level, or analyzed the word match rankings delivered by the relatedness technique with the rankings in the highest quality level. The last approach has gotten more gathering as it is less touchy to the genuine relatedness score esteems and considers a more sober minded correlation of the relatedness measures by and by. Such an approach guesses, to the point that keeping in mind the end goal to be viewed as an exact SR technique, the delivered rankings from the word combine orderings should be precise paying little respect to the real numerical esteem doled out to word sets. In any case, in the previous assessment strategy, the supreme SR esteems are thought to be vital with the support that the rankings in the highest quality level informational collections don't really precisely speak to the coveted word combine requesting. This is supported by the fact that in some cases, the gold standard word orderings are sensitive to very small difference between the word pair similarities and therefore, the correct order is questionable.

Table 1. Knowledge Resources

	Knowledge resource						
	Linguistically constructed			Collaboratively constructed		Web based	
System	WordNet	GermaNet	Other	Wikipedia	Wiktionary	Search Engine	Semantic Web
Resnik	✓						
Jiang and Conrath	✓						
Lesk	✓						
ESA						✓	
Cilibrasi and Vitányi				✓		✓	
WikiRelate!				✓			
Sahami and Heilman	✓						
Patwardhan and Pedersen	✓						
Hughes and Ramage				✓			✓
TSA				✓			
WLM							
Zesch et al					✓		
Gur		✓					
REWOrD							

Conclusion

In this survey paper, we study in details a comprehensive knowledge of SR methods, which reflected on various knowledge properties, methods and evaluations. First, we selected a representative set of SR approaches reported in the literature. We choose an agent set of SR approaches revealed in the writing. At that point, we to characterize these methodologies as indicated by the accompanying three measurement: information resource(s) utilized, the computational strategy connected for processing relatedness and the embraced assessment technique. By mapping the chose frameworks into the structure, we deliberately broke down the favorable circumstances and inconveniences of each distinguished information assets, relatedness computational technique, and assessment strategies. In this way, analysts who might need to additionally enhance or convey certain SR frameworks or strategies can exceptionally profit by the understanding gave by this examination.

References

[1] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M. & Soroa, A. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 19–27. Association for Computational Linguistics.

[2] Banerjee, S. & Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02), Gelbukh, A. F. (ed.). Springer-Verlag, 136–145.

[3] Bicici, M. E. 2015. RTM-DCU: predicting semantic similarity with referential translation machines. In SemEval-2015: Semantic Evaluation Exercises – International Workshop on Semantic Evaluation. http://doras. dcu.ie/20650/.

[4] Bollegala, D., Matsuo, Y. & Ishizuka, M. 2006. Disambiguating personal names on the web using automatically extracted key phrases. In Proceedings of the 17th European Conference on Artificial Intelligence, 553–557.IOS Press.

[5] Bollegala, D., Matsuo, Y. & Ishizuka, M. 2007. Measuring semantic similarity between words using web search engines. In Proceedings of the 16th International Conference on World Wide Web (WWW '07), 757–766.ACM.
[6] Bu, F., Hao, Y. & Zhu, X. 2011. Semantic relationship discovery with Wikipedia structure. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence – Vol. 3 (IJCAI '11), Walsh, T. (ed.). AAAI Press, 1770–1775.
[7] Budan, I. A. & Graeme, H. 2006. Evaluating WordNet-based measures of semantic distance. Computational Linguistics 32(1), 13–47.
[8] Budanitsky, A. & Hirst, G. 2006. Evaluating Wordnet-based measures of lexical semantic relatedness. Computational Linguistics 32(1), 13–47.
[9] Chen, H. H., Lin, M. S. & Wei, Y. C. 2006. Novel association measures using web search with double checking.In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, 1009–1016. Association for Computational Linguistics.
[10] Chen, P., Ding, W., Bowes, C. & Brown, D. 2009. A fully unsupervised word sense disambiguation method using dependency knowledge. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09), 28–36. Association for Computational Linguistics.
[11] Cilibrasi, R. L. & Vitanyi, P. 2007. The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383.
[12] Duan, J. & Zeng, J. 2012. Computing semantic relatedness based on search result analysis. In Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology – Vol. 3,205–209. IEEE Computer Society.
[13] Euzenat, J. & Shvaiko, P. 2013. Ontology Matching, 2nd edition. Springer-Verlag. Feng, Y., Fani, H., Bagheri, E. & Jovanovic, J. 2015. Lexical semantic relatedness for Twitter analytics. In IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI 2015), 202–209. IEEE.
[14] Gabrilovich, E. & Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07).Sangal, R., Mehta, H. & Bagga, R. K. (eds). Morgan Kaufmann Publishers Inc., 1606–1611.
[15] P. Resnik, (1995), "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, **(1995)** August 20-25; Montréal Québec, Canada.
[16] Lesk, Michael (1986). "Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." Proceedings of the 1986 SIGDOC Conference, Toronto, Canada, June 1986, 24-26.
[17] Miller, 1990 George Miller. "WordNet: An on-line lexical database". *International Journal of Lexicography*,3(4).
[18] Mahajan, S., Sharma, S., & Rana, V. (2017). Design a Perception Based Semantics Model for Knowledge Extraction. *International Journal of Computational Intelligence Research*, *13*(6), 1547-1556.
[19] Mahajan, S., & Rana, V. (2017). Spam Detection on Social Network through Sentiment Analysis. *Advances in Computational Sciences and Technology*, *10*(8), 2225-2231.
[20] Sunita, & Rana, V., Sharma, S.(2017). A Review: Phishing and its Impact. International Conference on recent innovations in science, Agriculture, Engineering and management. www.confereneceworld.in .