ENCYCLOPEDIA WITH SEMANTIC COMPUTING AND ROBOTIC INTELLIGENCE
Vol. 1, No. 1 (2017) 1630005 (5 pages)

DOI: 10.1142/S2425038416300056

World Scientific
www.worldscientific.com

Methods and resources for computing semantic relatedness

Yue Feng* and Ebrahim Bagheri†
Ryerson University
Toronto, Ontario, Canada
*yue.feng@ryerson.ca
†ebrahim.bagheri@ryerson.ca

Received 28 May 2016; Accepted 20 July 2016; Published 17 March 2017

Semantic relatedness (SR) is defined as a measurement that quantitatively identifies some form of lexical or functional association between two words or concepts based on the contextual or semantic similarity of those two words regardless of their syntactical differences. Section 1 of the entry outlines the working definition of SR and its applications and challenges. Section 2 identifies the knowledge resources that are popular among SR methods. Section 3 reviews the primary measurements used to calculate SR. Section 4 reviews the evaluation methodology which includes gold standard dataset and methods. Finally, Sec. 5 introduces further reading.

In order to develop appropriate SR methods, there are three key aspects that need to be examined: (1) the knowledge resources that are used as the source for extracting SR; (2) the methods that are used to quantify SR based on the adopted knowledge resource; and (3) the datasets and methods that are used for evaluating SR techniques. The first aspect involves the selection of knowledge bases such as WordNet or Wikipedia. Each knowledge base has its merits and downsides which can directly affect the accuracy and the coverage of the SR method. The second aspect relies on different methods for utilizing the beforehand selected knowledge resources, for example, methods that depend on the path between two words, or a vector representation of the word. As for the third aspect, the evaluation for SR methods consists of two aspects, namely (1) the datasets that are used and (2) the various performance measurement methods.

SR measures are increasingly applied in information retrieval to provide semantics between query and documents to reveal relatedness between non-syntactically-related content. Researchers have already applied many different information and knowledge sources in order to compute SR between two words. Empirical research has already shown that results of many of these SR techniques have reasonable correlation with human subjects interpretation of relatedness between two words.

Keywords: Semantic relatedness; information retrieval; similarity; natural language processing.

1. Overview of Semantic Relatedness

It is effortless for humans to determine the relatedness between two words based on the past experience that humans have in using and encountering related words in similar contexts. For example, as human beings, we know *car* and *drive* are highly related, while there is little connection between *car* and *notebook*. While the process of deciding semantic relatedness (SR) between two words is straightforward for humans, it is often challenging for machines to make a decision without having access to contextual knowledge surrounding each word. Formally, SR is defined as some form of lexical or functional association between two words rather than just lexical relations such as synonymy and hyponymy.[1]

1.1. *Applications*

Semantic relatedness is widely used in many practical applications, especially in natural language processing (NLP) such as word sense disambiguation,[2] information retrieval,[3] spelling correction[1] and document summarization, where it is used to quantify the relations between words or between words and documents.[4] SR is extremely useful in information retrieval techniques in terms of the retrieval process where it allows for the identification of semantic-related but lexically-dissimilar content.[1] Other more specialized domains such as biomedical informatics and geoinformatics have also taken advantages of SR techniques to measure the relationships between bioentities[5] and geographic concepts,[6] respectively.

1.2. *Challenges*

Developing SR methods is a formidable task which requires solutions for various challenges. Two primary challenges are encountered with the underlying knowledge resources and formalization of the relatedness measures respectively.

(1) Knowledge resources challenges: Knowledge resources provide descriptions for each word and its relations. Knowledge resources can be structured or unstructured, linguistically constructed by human subjects or collaboratively constructed through encyclopedia or web-based.

It is challenging to clean and process the large set of knowledge resources and represent each word with its extracted descriptions which requires considerable computation power.

(2) Formalization challenges: Designing algorithms to compute SR between words is also challenging since efficiency and accuracy are two important factors to be considered.

2. Knowledge Resources

In the world of SR techniques, the term knowledge resources refers to the source of information where the descriptions and relations of words are generated from. Five knowledge resources that are popular adopted literature are introduced below.

2.1. *WordNet*

WordNet is an English lexical database which is systematically developed by expert linguists. It is considered the most reliable knowledge resource due to the reason that it has been curated through a well-reviewed and controlled process. WordNet provides descriptions for English words and expresses various meanings for a word which is polysemy according to different contexts. Expert linguists defined relations and synsets in WordNet which are two of the main parts where the relations express the relations between two or more words such as hypernymy, antonymy and hyponymy, and synsets are a set of synonymous words. Moreover, a short piece of text called gloss is attached to describe members of each synset.

WordNet has been widely applied in researches for computing the degree of SR. For example, Rada *et al.*[7] constructed a word graph whose nodes are Wordnet synsets and edges are associated relations. Then SR is represented as the shortest path between two nodes. Glosses defined in Wordnet have also been explored to compute SR. For instance, Lesk[8] introduced his method in 1986 that is counting the word overlap between two glosses where the higher count of overlap indicates higher SR between the two words.

A German version of Wordnet has also been constructed named GermaNet. GermaNet shares all the features from Wordnet except it does not include glosses, therefore, approaches based on glosses are not directly applicable on GermaNet. However, Gurevych[9] has proposed an approach to solve the problem by generating pseudo-glosses for a target word where the pseudo-glosses are the set of words that are in close relations to the target word in the relationship hierarchy.

2.2. *Wikipedia*

Wikipedia provides peer-review and content moderation processes to ensure reliable information. The information in Wikipedia is presented as a collection of articles where each article is focused on one specific concept. Besides articles, Wikipedia contains hyperlinks between articles, categories and disambiguation pages.

Some researchers have benefited from the textual content of Wikipedia articles. For example, a widely-used SR technique called explicit semantic analysis (ESA)[10] treats a target word as a concept and uses its corresponding Wikipedia article as the knowledge resource to describe the target word; therefore, each word is represented as a vector of words from the associated Wikipedia article and the weights are the TF-IDF values of the words. Then cosine similarity method is applied on two vectors for two words respectively to calculate SR. Besides exploring the article contents, hyperlinks between Wikipedia articles can also be used to establish relationships between two words. Milne and Witten[11] and Milne[12] represented each word as a weighted vector of links obtained through the number of links on the corresponding Wikipedia article and the probability of the links occurrences. In their work, they have proved that processing only links on Wikipedia is more efficient and can achieve comparable results with ESA. The Wikipedia category system has also been exploited for the task of SR. For instance, WikiRelate[13] expressed the idea that SR between two words is dependent on the relatedness of their categories, therefore, they represented each word with their related category.

2.3. *Wiktionary*

Wiktionary is designed as a lexical companion to Wikipedia which is a multilingual, Web-based dictionary. Similar to Wordnet, Wiktionary includes words, lexical relations between words and glosses. Researchers have taken advantages of the large number of words in Wiktionary to create high dimensional concept vectors. For example, Zesch *et al.*[14] constructed a concept vector for each word where the value of the term is the TF-IDF score in the corresponding Wiktionary entry. Then the SR is calculated based on the cosine similarity of the two concept vectors. Also, given the fact that Wiktionary consists of lexical-semantic relations embedded in the structure of each Wiktionary entry, researchers have also considered Wiktionary as a knowledge resource for computing SR. For instance, Krizhanovsky and Lin[15] built a graph from Wiktionary where nodes are the words and the edges are the lexical-semantic relations between pairs of words. Then they applied path-based method on the graph to find SR between words. Similar to Wordnet, the glosses provided by Wiktionary are explored. Meyer and Gurevych[16] performed sense disambiguation process based on word overlaps between glosses.

2.4. *Web search engines*

Given Web search engines provide access to over 45 billion web pages on the World Wide Web, their results have been used as a knowledge source for SR. For a given search query, search engines will return a collection of useful information including rich snippets that are short pieces of text each containing a set of terms describing the result page, Web page URIs, user-specified metadata and descriptive page titles. Works based on search engines snippets include the method from Spanakis *et al.*[17] in which they extracted lexico-synactic patterns from snippets with the assumption that related words should have similar patterns. Duan and Zeng[18] computed the SR based on the co-occurrences of the two words and occurrences of each word from the snippets returned by the search engine. Also there are some works that rely on the content of the retrieved pages. For example, Sahami and Heilman[19] enhanced the snippets by including the top-*k* words with the highest TF-IDF value from each of the returned page to represent a target word.

2.5. *Semantic web*

Some researchers have exploited the Semantic Web and the Web of Data. The data on the Web of Data is structured so that it can be interlinked. Also, the collection of Semantic Web technologies such as RDF, and OWL among others allows for running queries. REWOrD[20] is one of the earlier works in this area. In this work, each target word is represented as a vector where each element is generated from RDF predicates and their informativeness scores. The predicates are obtained from DBpedia triples where they correspond to each word and the informativenss scores are computed based on predicate frequency and inverse triple frequency. After that, the cosine similarity method is applied on the vectors to generate the SR between two words. The semantic relations defined by the Web ontology language (OWL) have also been explored, for example, In Karanastasi and Christodoulakiss model,[21] three facts that are (1) the number of common properties and the inverseOf properties that the two concepts share; (2) the path distance between two concepts common subsumer; and (3) the count of the common nouns and synonyms from the concepts description are combined to compute SR.

3. Semantic Relatedness Methods

Many SR methods have been developed by manipulating the information extracted from the selected knowledge resources. Some methods use the relationships between each word from the knowledge resource to create a graph and apply these relations to indicate SR, while other methods directly use content provided by the knowledge resource to represent each concept as a vector and apply vector similarity methods to compute the SR. Moreover, there have been works on temporal modeling for building SR techniques.

3.1. *Resnik*

Resnik[22] proposed his model in 1995. The idea is that the more information two words share, the higher their SR will be. Therefore, the IS-A hierarchy is adopted to find the lowest common subsumer of two words in a taxonomy, then the information content value is calculated as the SR score.

3.2. *WikiRelate!*

Strube and Ponzetto[13] created a graph based on the information extracted from Wikipedia where nodes are Wikipedia articles, and the edges are the links between the articles. Then the shortest path is selected between two words which are Wikipedia articles to determine the SR score.

3.3. *Hughes and Ramage*

Hughes and Ramage[23] construct a graph from WordNet where the nodes are Synsets, TokenPOS and Tokens, and the edges are the relations defined in WordNet between these nodes. The conditional probability from one node to another is caluclated beforehand, then the authors apply Random Walk algorithm on the graph to create a stationary distribution for each target word by starting the walk on the target word node. Finally, SR is computed by comparing the similarity between the stationary distributions obtained for two words.

3.4. *ESA*

Gabrilovich adn Markovitch[10] have proposed the ESA technique in 2007 by considering Wikipedia as its knowledge resource. In their approach, a semantic mapper is built to represent a target word as a vector of Wikipedia concepts where the weights are the TF-IDF values of the words in the underlying articles. Then the SR is computed by calculating the similarity between two vectors represented for the two words respectively.

3.5. *Lesk*

Lesk[8] takes advantage of the glosses defined for each word from WordNet. Specifically, SR is determined by counting the number of words overlap between two glosses obtained for the two words. The higher the count of overlap, the more related the two words are.

3.6. *Sahami and Heilman*

Sahami and Heilman[19] benefit from the results returned by a Web search engine. By querying the target word, they enrich the short snippets by including the top words ranked based on the TF-IDF values from each returned page. Then the vector is used to compute the degree of SR between two words.

3.7. *WLM*

Milne[11] intends to reduce the computation costs of the ESA approach, therefore, a more efficient model is built by considering links found within corresponding Wikipedia articles where the basic assumption is the more links two articles share, the more they are related. So a word is represented as a vector of links. Finally, SR is computed by comparing the similarity between the link vectors.

3.8. *TSA*

Radinsky *et al.*[24] propose a temporal semantic analysis method based on the idea that enormous information can be revealed by studying the similarity of word usage patterns over time. Therefore, in their model, a word is represented as a weighted vector of concept time series obtained from a historical archive such as NY Times archive. Then SR is found by comparing the similarity between two time series.

4. Evaluation

In order to evaluate a SR method, researchers have adopted various goldstandard datasets and strategies for comparative analysis. In this section, we introduce the common datasets and metrics researchers have used.

4.1. *Datasets*

The gold standard datasets are often constructed by collecting subjective opinion of humans in terms of the SR between words. The main purpose of creating a SR dataset is to assign a degree of SR between a set of word pairs so they can be used as a gold standard benchmark for evaluating different SR methods. The datasets that have been used and cited in literatures are mainly in English and German languages. Below are four popular English datasets.

4.1.1. *RG-65*

The Rubenstein–Goodenough (RG-65)[25] is created by collecting human judgments from 51 subjects, the similarity between each word pair is equal to the average of the scores given by the subjects. The RG-65 dataset includes 65 noun pairs, and the similarity of each word pair is scored on a scale between 0 to 4 where higher score indicates higher similarity. The RG-65 dataset has been used as gold standard in many researches such as Strube and Ponzetto.[13]

4.1.2. *MC-30*

Miller–Charles (MC-30)[26] is a subset of the original RG-65 dataset that contains 30 noun pairs. The MC-30 dataset is additionally verified and evaluated by another 38 subjects and it is widely adopted in many works such as in Refs. 11 and 17.

4.1.3. *Fin-353*

Finkelstein *et al.*[27] introduced a dataset that contains 353 word pairs where 30 word pairs are obtained from the MC-30 dataset. The dataset is divided into two parts where the first part contains 153 word pairs obtained from 13 subjects and the second part contains 200 word pairs that are judged from 16 subjects. In some literature, the first set is used for training and the second is used for evaluation. The use of Fin-353 dataset can be found in Ref. 28 among others.

4.1.4. *YP-130*

Yang Powers (YP-130) is a dataset designed especially for evaluating a SR methods ability to assign the relatedness between verbs. The YP-130 contains 130 verb pairs.

There are also some datasets in German language. For instance, Gurevych dataset (Gur-65)[9] is the German translation of the English RG-65 dataset, Gurevych dataset (Gur-30) is a subset of the Gur-65 dataset, which is associated with the English MC-30 dataset. Gurevych dataset (Gur-350)[29] consists of 350 word pairs which includes nouns, verbs and adjectives judged by eight human subjects. The Zesch–Gurvych (ZG-222) dataset[29] contains 222 domain specific word pairs that were evaluated by 21 subjects which includes nouns, verbs and adjectives.

4.2. *Methods*

There are two typical ways to evaluate a SR method that are (1) calculating the degree of correlation with human judgments and (2) measuring performance in application-specific tasks.

4.2.1. *Correlation with human judgments*

Calculating the correlation between the output of a SR method and the score obtained from a gold standard dataset is one of the main techniques for evaluating a semantic method. Either the absolute values from a semantic method and the relatedness values from the gold standard are used, or the rankings produced by the relatedness method with the rankings in the gold standard are compared. Comparing the correlation between rankings is more popularly adopted in literature due to the reason it is less sensitive to the actual relatedness values. Pearson product-moment correlation coefficient[31] and Spearmans rank correlation coefficient[30] are two most popular coefficient to calculate the correlation between a SR method and the human judgments.

4.2.2. *Application-specific tasks*

Instead of directly comparing the output from a SR method with the gold standard dataset, a SR method can be embedded into an application-specific task, and the performance of the application can be the indicator of the performance of the SR method. The underlying hypothesis of this evaluation is that the more accurate a SR method is, the better the performance of the application task.

Various application-specific tasks have been used to evaluate the SR method. For instance, Sahami and Heilman[19] evaluated their work through the task of search query suggestion; Patwardhan and Pedersen[2] used their SR method in the word sense disambiguation application as the target evaluation application; while Gracia and Mena[32] deployed their method in the ontology matching task.

References

[1] I. A. Budan and H. Graeme, Evaluating wordnet-based measures of semantic distance, *Comput. Linguist.* **32**(1), 13 (2006).

[2] S. Patwardhan, S. Banerjee and T. Pedersen, SenseRelate: TargetWord: A generalized framework for word sense disambiguation, *Proc. ACL 2005 on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics* (2005).

[3] L. Finkelstein *et al.*, Placing search in context: The concept revisited, *Proc. 10th Int. Conf. World Wide Web*, ACM (2001).

[4] C. W. Leong and R. Mihalcea, Measuring the semantic relatedness between words and images, *Proc. Ninth Int. Conf. Computational Semantics, Association for Computational Linguistics* (2011).

[5] M. E. Renda *et al.*, *Information Technology in Bio- and Medical Informatics*.

[6] B. Hecht *et al.*, Explanatory semantic relatedness and explicit spatialization for exploratory search, *Proc. 35th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, ACM (2012).

[7] R. Rada *et al.*, Development and application of a metric on semantic nets, *IEEE Trans. Syst. Man Cybern.* **19**(1), 17 (1989).

[8] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, *Proc. 5th Ann. Int. Conf. Systems Documentation*, ACM (1986).

[9] I. Gurevych, Using the structure of a conceptual network in computing semantic relatedness, *Natural Language Processing IJCNLP 2005* (Springer Berlin Heidelberg, 2005), pp. 767–778.

[10] E. Gabrilovich and S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, *IJCAI*, Vol. 7 (2007).

[11] I. Witten and D. Milne, An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, *Proc. AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, AAAI Press, Chicago, USA (2008).

[12] D. Milne, Computing semantic relatedness using wikipedia link structure, *Proc. New Zealand Computer Science Research Student Conference* (2007).

[13] M. Strube and S. P. Ponzetto, WikiRelate! Computing semantic relatedness using Wikipedia, *AAAI*, Vol. 6 (2006).

[14] T. Zesch, C. Mller and I. Gurevych, Using wiktionary for computing semantic relatedness, AAAI, Vol. 8.

[15] A. A. Krizhanovsky and L. Feiyu, Related terms search based on wordnet/wiktionary and its application in ontology matching, arXiv:0907.2209.

[16] C. M. Meyer and I. Gurevych, To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity, COLING 2012 (2008).

[17] G. Spanakis, G. Siolas and A. Stafylopatis, A hybrid web-based measure for computing semantic relatedness between words, *21st Int. Conf. Tools with Artificial Intelligence, 2009. ICTAI'09*, IEEE (2009).

[18] J. Duan and J. Zeng, Computing semantic relatedness based on search result analysis, *Proc. The 2012 IEEE/WIC/ACM Int. Joint Conf. Web Intelligence and Intelligent Agent Technology* Vol. 3, IEEE Computer Society (2012).

[19] M. Sahami and T. D. Heilman, A web-based kernel function for measuring the similarity of short text snippets, *Proc. 15th Int. Conf. World Wide Web*, ACM (2006).

[20] G. Pirr, REWOrD: Semantic relatedness in the web of data, *AAAI* (2012).

[21] A. Karanastasi and S. Christodoulakis, The OntoNL semantic relatedness measure for OWL ontologies, *2nd Int. Conf. Digital Information Management, 2007. ICDIM'07*, Vol. 1. IEEE (2007).

[22] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, arXiv:cmp-lg/9511007.

[23] T. Hughes and D. Ramage, Lexical Semantic Relatedness with Random Graph Walks. EMNLP-CoNLL (2007).

[24] K. Radinsky *et al.*, A word at a time: Computing word relatedness using temporal semantic analysis, *Proc. 20th Int. Conf. World Wide Web*, ACM (2011).

[25] H. Rubenstein and J. B. Goodenough, Contextual correlates of synonymy, *Commun. ACM* **8**(10), 627 (1965).

[26] G. A. Miller and W. G. Charles, Contextual correlates of semantic similarity, *Lang. Cogn. Process.* **6**(1), 1 (1991).

[27] E. Agirre *et al.*, A study on similarity and relatedness using distributional and wordnet-based approaches, *Proc. Human Language Technologies: The 2009 Annual Conf. North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics* (2009).

[28] C. Fellbaum, *WordNet* (Blackwell Publishing Ltd, 1998).

[29] T. Zesch and I. Gurevych, Automatically creating datasets for measures of semantic relatedness, *Proc. Workshop on Linguistic Distances, Association for Computational Linguistics* (2006).

[30] J. H. Zar, Spearman rank correlation, *Encyclopedia of Biostatistics* (1998).

[31] J. Benesty *et al.*, Pearson correlation coefficient, *Noise Reduction in Speech Processing* (Springer Berlin Heidelberg, 2009), pp. 1–4.

[32] J. Gracia and E. Mena, Web-based measure of semantic relatedness, *Web Information Systems Engineering-WISE 2008* (Springer Berlin Heidelberg, 2008), pp. 136–150.