Short text similarity measurement methods: a review

Prakoso, D.W.; Abdi, A.; Amrit, C.

DOI
10.1007/s00500-020-05479-2
Publication date
2021
Document Version
Final published version
Published in
Soft Computing

Soft Computing (2021) 25:4699–4723
https://doi.org/10.1007/s00500-020-05479-2

METHODOLOGIES AND APPLICATION

Short text similarity measurement methods: a review

Dimas Wibisono Prakoso[1] · Asad Abdi[1] · Chintan Amrit[2]

Published online: 3 January 2021

Abstract
Short text similarity measurement methods play an important role in many applications within natural language processing. This paper reviews the research literature on short text similarity (STS) measurement method with the aim to (i) classify and give a broad overview of existing techniques; (ii) find out its strengths and weaknesses in terms of the domain the independence, language independence, requirement of semantic knowledge, corpus and training data, ability to identify semantic meaning, word order similarity and polysemy; and (iii) identify semantic knowledge and corpus resource that can be utilized for the STS measurement methods. Furthermore, our study also considers various issues such as the difference between the various text similarity methods and the difference between semantic knowledge sources and corpora for text similarity. Although there are a few review papers in this area, they focus mostly only on one/two existing techniques. Furthermore, existing review papers do not cover recent research. To the best of our knowledge, this is a comprehensive systematic literature review on this topic. The findings of this research can be as follows: It identified four semantic knowledge and eight corpus resources as external resources that can be classified into general-purpose and domain-specific. Furthermore, the existing techniques can be classified into string-based, corpus-based, knowledge-based and hybrid-based. Moreover, expert researchers can utilize this review as a benchmark as well as reference to the limitations of current techniques. The paper also identifies the open issues that can be considered as feasible opportunities for future research directions.

Keywords Natural language processing · Text mining · Linguistics · Similarity measures · Semantics · Syntax

1 Introduction

The short text similarity is a topic that has been studied in the field of computer science, especially language processing. It plays an important role in many applications within natural language processing (NLP) and related areas such as question and answering systems (Aouicha et al. 2018), a conversational agent in the business, gene clustering in biomedical, text summarization (Alguliyev et al. 2017), web page, web image retrieval and plagiarism detection (Abdi et al. 2017), essay scoring, information retrieval, text classification and text clustering (Abualigah et al. 2017, 2018a, b; Abualigah 2019). On the other hand, with the express growth of online social network, users have joined these networks. In these digital worlds, users present themselves, share information about their favorites, interest and behavior or share their personal opinion on some issues of economic, social, cultural, etc. Consequently, through several activities on social network such as posting entries, sharing video clips, images, comment and like, huge data are created on the social network. These huge data attract many researchers, businessmen, etc. to mine and exploit it. It also brings some new challenges to researchers. For instance, the basic issues in these challenges are the problem of estimating the similarity among users on social networks based on their profile, interest and comments.

Communicated by V. Loia.

✉ Asad Abdi
s.abdiesfandani@utwente.nl

Dimas Wibisono Prakoso
dimaswibisonoprakoso@student.utwente.nl

Chintan Amrit
c.amrit@uva.nl

[1] Department of Industrial Engineering and Business Information Systems, University of Twente, Enschede, The Netherlands

[2] Department of Operations Management, Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands

Short text similarity (STS) measurement aims to determine the degree of similarity between pairs of short texts. The similarity itself is not only examined from the lexical point of view which only considers the sequences of characters but also should take into account the semantic meaning. Two short texts might be composed of different words, but semantically would be similar. For instance, Table 1 shows a pair of sentences and human-labeled degree of similarity.

The STS has attracted lots of attention as a hot research topic. The researchers have also proposed many methods to the problem of measuring STS since it lies at the core of many applications in natural language processing to measure the short text similarity. However, each method has its strengths and weaknesses, and these methods are usually limited to handling a specific problem. Thus, a review of the related issues and collecting comprehensive information is needed as few papers have been published in the area of the STS. However, existing survey papers do not contain recently published papers. Furthermore, unlike other papers, our study considers various issues such as (i) the recent text similarity methods that have been proposed and their classification, (ii) comparison between the different text similarity methods and (iii) the difference between semantic knowledge sources and corpora for text similarity.

The current survey paper also prepares a structured and broad overview of the existing methods and then categorizes them into string-based, corpus-based, knowledge-based and hybrid-based methods. Furthermore, it provides a global view of most recent results in the literature that are not part of the existing surveys. Moreover, the survey also considers the characteristics, weaknesses and strengths of proposed methods to select an effective and suitable approach for a specific problem. This discusses the future direction to improve the performances of existing methods. Despite the advancement of information technology, there are still some outstanding issues in this research area. Thus, it is necessary to examine the progress, prepare a comprehensive survey of the current situation of the research field, highlight the advances obtained and indicate remaining drawbacks. This study not only complements earlier review papers but also includes much additional information.

However, to achieve our aim, we formulate our research questions as follows: RQ1: what are the text similarity methods that have been reported in the literature?; RQ2: how do the different text similarity methods compare with each other as reported in the literature?; RQ3: what are the different semantic knowledge sources and corpora for text similarity? This review was conducted based on Kitchenham (2004) using snowballing techniques from Wohlin (2014) for performing systematic literature reviews. A comprehensive search for relevant literature was conducted via seven electronic database resources.

The remaining sections of the paper are organized as follows. The summary of the related works is discussed in Sect. 2. We will describe the methodology of this study in Sect. 3. Section 4 presents the result of the research based on the literature review. In this section, the research questions are answered and explained in detail. In Sect. 5, the result of the SLR is discussed. The significant implications of the current research are presented in Sect. 6. Finally, the main conclusion and a view for developing future work are drawn in Sect. 7.

2 Related work

While conducting the literature review, we also encountered other reviews that are closely related to STS.

Gomaa and Fahmy (2013) published a survey paper. The paper presents a survey on methods to measure text similarity. The paper includes 38 methods in total and classified them into four categories which are string-based, corpus-based, knowledge-based and hybrid-based similarity. String-based methods measure similarity by considering string sequence of texts. The corpus-based method calculates similarity by utilizing large corpora to determine the relationship between words that compose the texts. On the other hand, knowledge-based methods use information derived from semantic networks. The last method, hybrid, joins several methods to aggregate the advantages of each method used. The current paper includes the methods that have been proposed more than 5 years ago.

However, there can be much progress in the field of STS that are not covered by this study. It is a gap that the current SLR aims to fill. Other studies focus on only several methods of text similarity.

Kaundal and Kaur (2017) published a review of measuring short text semantic similarity (STSS) by using two techniques which are vector space model and knowledge-

Table 1 Human-assigned similarity score to sentence

No.	Sentence 1	Sentence 2	Similarity (%)
1.	I like that bachelor	I like that unmarried man.	80
2.	Red alcoholic drink	A bottle of wine.	50
3.	I have a pen	Where do you live?	0

based model by incorporating WordNet. Altszyler et al. (2016) focus on the comparison between LSA and word2vec using small corpora. However, our systematic literature review focused on text similarity with wider coverage of papers.

Furthermore, the SLRs on semantic similarity measures that were conducted by Elavarasi et al. (2014) and Majumder et al. (2016) do not cover neural network-based method (word embedding techniques) such as word2vec (Mikolov et al. 2013). Recently, neural network language embedding has gained popularity and is considered as a breakthrough in language processing technology (Goth 2016). However, our SLR aims at contributing not only to the knowledge of researchers but also to the stakeholders in text similarity practice.

2.1 Similarity measure for social networks

For many people, day-to-day interaction has been replaced by instant messages, likes, share and retweet through social networking websites. Social networking sites are increasing at the alarming rate. Online social networking has become popular among all the age groups and it has tied the people all over the world. It does not matter to which group a person belongs to or in which field the person is working, i.e., researchers, industrialists, celebrities, government officials, politicians, entrepreneurs, sportsperson, etc. Social networks have connected all of them.

Social network indicates a particular domain as a collection of nodes or profiles and links between them. In other words, a social network can be created from relational data and can be defined as a set of social entities, such as people, groups and organizations, with some relationships or interactions between them. It was born to enable users to share, express, interact and cultivate relationships on social and professional level. Due to its potential, significant scientific and technological efforts are created to better understand, control and extend this phenomenon. The public accessibility of web-based social networks stimulated extensive research in this domain. Understanding how networks grow and change, and being able to predict their behavior, contributes to the evolution of other domains such as business, education, social, biology, fraud detection and criminal investigation. Figure 1 presents the connectivity of users on the social network. The users from various countries interact with each other using social networking sites.

There are different forms of social networks[1] sites such as (Goyal and Singh 2016; Tabassum et al. 2018) friendship networks (Facebook, MySpace, etc.), follower networks (Twitter, LinkedIn, Pinterest, etc.), blogs (Blogger, LiveJournal), wiki sites (Wikipedia, Wetpaint), interaction networks (Emails, Whatsapp, Snapchat, etc.), multimedia sharing (Flickr, Youtube), user citation networks (Dblp, Science direct, Wikibooks, etc.) and Instagram.

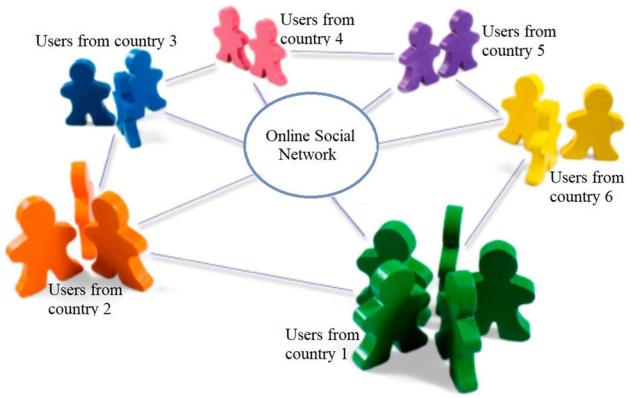


Fig. 1 Interaction among users using social networking website

Representation of social network a social network is a social structure that includes a finite set of social actors and their relationships. An actor is the social entity who participates in a certain network and who is able to act and form connections with other actors. It could be an individual or a organization. A social network is usually modeled by graph (Díaz and Ralescu 2012; Pupazan and Bhulai 2011), where vertices represent the social entities and edges represent the ties established between them. In graph terminology, vertices are also known as nodes refer to actors or subjects. Edges, also known as links or ties, describe the relationship between actors. However, a social network is a graph G that consists of two main components (Cordeiro et al. 2018): vertices (V) and a set of edges (E). It can be defined as $G = (V, E)$. Vertices represent objects, states, positions, placeholders and are represented by a set of unique vertices, where V can be demonstrated by $\{v_1, v_2, v_3, \ldots, v_n\}$. For each edge e ∈ E, there is associated a pair of graph vertices m, n. It can be formulated as $\forall_{e \in E} e \rightarrow (v_1, v_2)$, where v_1, $v_2 \in V$. Edges can be directed or undirected and can be weighted (or labeled) or unweighted. An undirected edge $e = (v_1, v_2)$ with v_1, v_2 ∈V, indicates that the relationship is bidirectional, that is, can go from v_1 to v_2 and vice versa. A directed edge $e = (v_1, v_2)$ specifies a one-directional relationship, that is, can only go from v_1 to v_2; this means that $(v_1, v_2) \neq (v_2, v_1)$. The total number of vertices n of graph G, $|V| = n$, is called the graph order. The total number of edges $|E| = m$ is known as the size of the graph G. The maximum number of edges in a undirected graph is $m_{\max} = \frac{n(n-1)}{2}$, while for the directed ones, it is $m_{\max} = n(n-1)$.

Similarity measure between nodes nowadays the data generated from many of the real-world applications are represented as a network of interconnected objects. In other

[1] http://www.ebizmba.com/articles/social-networking-websites.

words, besides the relations, a social network often represents flow of information, interactions and similarities, among the set of social actors (Rawashdeh and Ralescu 2015; Tabassum et al. 2018). The main objective is to extract more information and tackle different problems for various purposes; for example, actors may have multiple profiles over multiple networks or on the same network. Therefore, duplicate user profiles can be detected using the various similarity measures like cosine similarity measure. However, social network analysis (SNA) is a technique that can be use to tackle the aforementioned problems. SNA focuses on the relationships established between social actors to examine relationships and understand the relations among actors in social networks. This task is extremely useful in the process of extracting knowledge from networks and in the process of any problem-solving. Due to the high potential opened by this kind of analyses, SNA is used in various domains and fields such as terrorist networking, political and economic systems, educational systems and many others. Usually, an entry on a social network can be a video, an image, a text or a combination of all content. The problem of identifying the similarity or the difference between users, nodes or actors is not only based on the user profile on the social network, but also based on the data about user such as posting entries and commenting. However, a problem is to consider and calculate the similarity of users based on the entry that focuses on estimating similarity between textual content. This problem has been attracting many researchers. Hence, the methods presented in this study can be useful and used to estimate the similarity among textual values.

Summing up, this review paper makes a significant contribution compared to the aforementioned papers due to the following reasons: (i) The present study is one of the first efforts to focus on all existing method for short text similarity measure. (ii) Our survey contains recently published papers, while the existing papers do not include recently published papers. (iii) Our review paper indicates several factors that affect the punctuality of text similarity measure. (iv) Finally, unlike other surveys, our work explains the strengths and limitations of the proposed methods.

3 Research method

This segment presents the process involved in conducting the SLR. According to the previous study (Kitchenham 2004), the systematic review can be defined as the process of identifying as well as interpreting all available research with the target of answering a well-defined research question in a given topic area or phenomenon of interest. SLR introduces a more systematic approach to synthesize the research through the use of inclusion as well as exclusion criteria that provide the borders of evidence to be incorporated in the review. We follow SLR guideline provided by Kitchenham (2004) to identify gaps in the existing research and draw conclusions based on our research questions. In other words, by following an SLR guideline, a researcher will have a clear set of procedures to follow in reviewing the material for research and to identify where this material could support or conflict with their work (Budgen and Brereton 2006).The SLR guideline by Kitchenham (2004) is the de facto standard for literature review in the software engineering field. The guideline mainly comprises of three phases which are planning, conducting and reporting as summarized in Fig. 2.

In the planning phase, the aim of this review was clearly identified in conjunction with the following events: We discovered that there was no systematic review that covers neural network-based methods, such as word embedding techniques. Most of the existing SLRs were carried out a fairly long time ago. Furthermore, the previous studies (SLRs) did not focus on text similarity with a wider coverage of papers. Hence, we recognize the need for conducting this review based on the results from the past studies that covered text similarity. As an essential activity of the current review, we also derived the required research questions as discussed in Introduction.

A review protocol specifies the methods that will be used to undertake a specific systematic review. It comprises the definition of rational of the survey, research questions, search strategy, study selection criteria and procedure, study quality assessment, data extraction strategy and data synthesis. A review protocol is necessary to prevent researcher bias where a selection of studies or analyses is driven by researcher expectation. After a review protocol is defined, the conducting phase is executed by following the review protocol as described in Fig. 3.

In the study selection part of the conducting phase (Fig. 1), we combine the step with the snowballing approach based on guidance by Wohlin (2014) as illustrated in figure below. After a primary study is defined, we conduct forward and backward snowballing to expand the coverage of the literature search. The expansion might find the literature that is also relevant to the research questions.

3.1 Search strategy and resource database

In order to answer the aforementioned research questions, we selected digital databases that include the majority of journals as well as conference papers published within the computer science field in order to discover relevant studies for the review. We also set the beginning and end date for our review, since a beginning as well as a end date for a

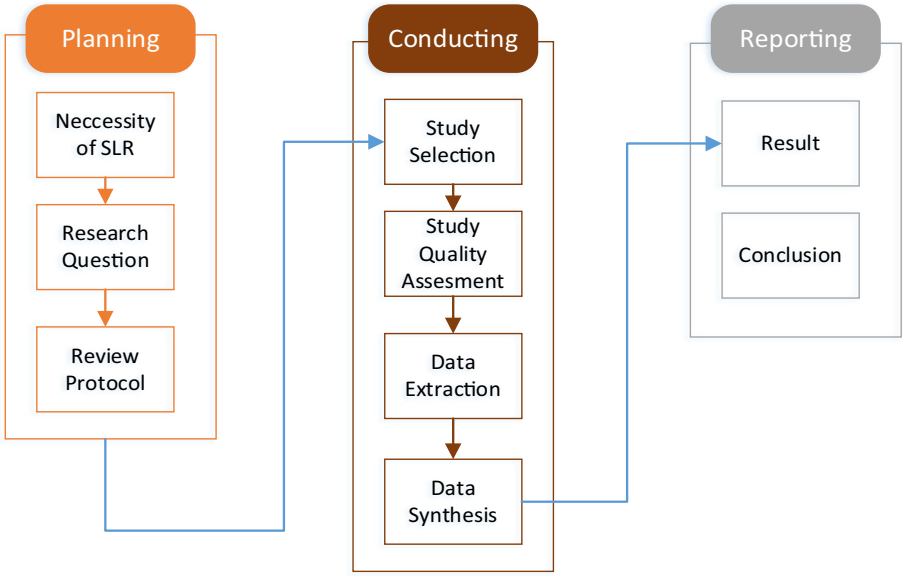


Fig. 2 Systematic literature review by Kitchenham (2004)

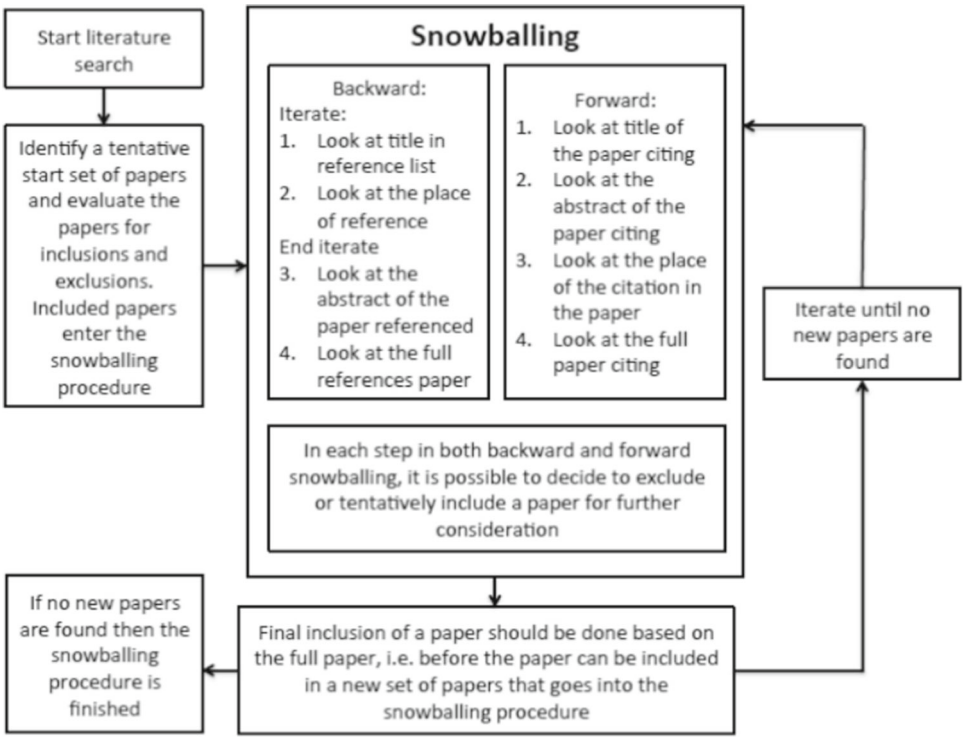


Fig. 3 Snowballing procedure by Wohlin (2014)

review is one of the strategies of a systematic review according to the study by Stapić et al. (2012).

The searches were narrowed to journal as well as conference proceedings that were published within the predetermined period. We used several terms for the STS, such as short text, text, sentence, similarity, measure, syntactic, lexical, semantic, corpus, semantic net and knowledge. Similarly, various terms and synonyms have been used for STS methods, including approaches, techniques and algorithms. In the current SLR, we targeted those databases that are considered important for an academic-oriented domain and used by students, researchers, as well as other scholars. Therefore, in order to capture the relevant papers, we conducted the searches on online databases using a search string. To create the search string, we use the Boolean OR to include substitutes and alternative words. The Boolean AND was used to link the foremost terms from population, intervention and context. Hence, the comprehensive search string is derived as:

((“*short text*” OR “*text*” OR “*sentence*”) AND (“*similarity*”) AND (“*syntactic*” OR “*lexical*” OR “*semantic*”) AND (“*corpus*” OR “*semantic net*” OR “*knowledge*”)AND (“*measure*” OR “*measurement*”) AND (“*approach*” OR “*technique*” OR “*method*” OR “*algorithm*”)).

In total, we have explored through several databases that comprised computer science-related articles and retrieved a total of 3398 papers.

3.2 Study selection

The papers were reviewed by going through the abstract, the introduction then the conclusion. In the initial grouping, the inclusion and the exclusion criteria were also applied to remove irrelevant studies according to the screening of titles as well as the abstracts. When the titles, abstracts and conclusions were not enough to determine the relevance of the paper, we then referred to the full text. We used the following inclusion and exclusion criteria to select the relevant studies.

Inclusion criteria:

1. The study is peer-reviewed.
2. It has the answer to at least one of the research questions
3. It is relevant to the search terms defined in Sect. 3.1
4. The study includes a detailed empirical evaluation.
5. If more than one paper reports the same study, only the latest or fullest paper was included.

Exclusion criteria:

1. Abstract papers with no full text available are excluded.
2. The language of the paper is not English.
3. Short papers with less than four pages.
4. Duplicated studies (by title or content)

At the end of the study selection process where primary studies have been identified, the forward and backward snowballing method by Wohlin (2014) is applied to extend the coverage of the search result. The overall selection phases is summarized in Fig. 4.

Primary search using string search produced 3398 studies. The number of studies was then significantly reduced in the secondary search stage which examined the title, abstract and conclusion. Then we applied inclusion and exclusion criteria so that the potential primary study was reduced further to 29 papers. Backward and forward snowballing were applied to references, resulting in 6 additional studies. In total, the study selection process produced 35 primary studies.

3.3 Study quality assessment

Additionally, in the process of study selection, we also specified the following quality assessment criteria so that the SLR could produce reliable and high-quality result and conclusion.

- Criteria 1: Study contribution is clearly described.
- Criteria 2: Artifacts and methods used in the study are clearly described.
- Criteria 3: Empirical validation is performed.
- Criteria 4: The results and applications are described and discussed thoroughly.

3.4 Data extraction and synthesis

Extraction of the data plan is structured to precisely record the facts acquired by the researchers after the primary studies. Considering the above-mentioned criteria, 35 articles were carefully chosen for our review in order to reveal answers to the research questions identified. Additionally, we also extracted data to compile bibliographic information. The types of data we extract from our paper are summarized in Table 2.

Regarding publication time, Fig. 5 shows the distribution of 35 primary studies per year. In Fig. 4, we could see several papers were published before 2006. The papers were about the classic method of STS measurement, which only compares sequences of characters or words without considering the semantic meaning of the sentence.

For the following years, the publication of papers in this field was relatively stable, except in 2012 and 2013. On these years, there was a significant increase due to the existence of the SemEval 2012 conference. At this conference, there was one competition named semantic text similarity where 88 methods were submitted (Agirre et al. 2012). However, for this SLR, we only reviewed methods that were ranked in the top 3.

4 Analysis and discussion

In the following sections, we will answer the predefined research questions. However, we selected digital databases that include the majority of journals as well as conference papers published within the computer science field to discover relevant studies for the review.

4.1 Answer to RQ1

The STS methods identified by our literature research can be categorized into string-based, corpus-based, knowledge-

Fig. 4 Study selection process

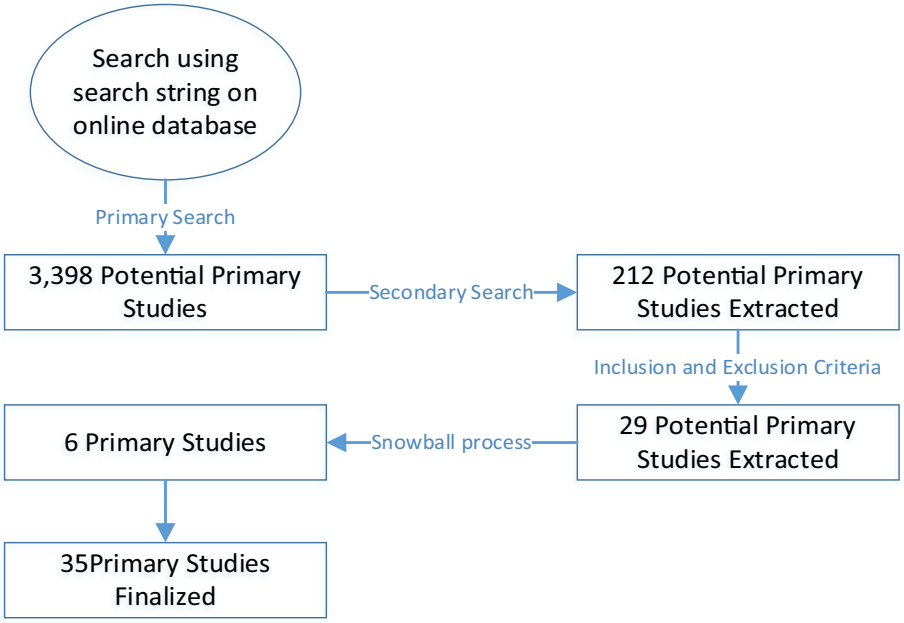
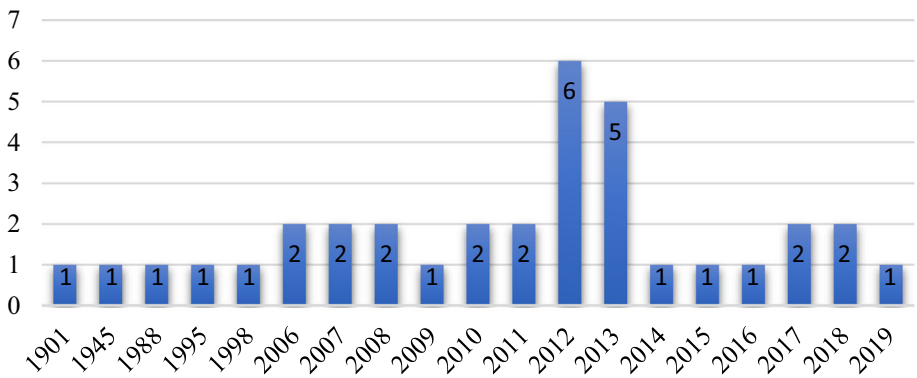


Table 2 Data extracted from the paper

Type of the data	Description
Study ID	Unique ID for each paper
Year	The year when the paper was published
Author	The author of the paper
Title	The title of the paper
Venue	Publication venue of the research, e.g., conference proceeding, journal
Technique	Characteristics and techniques used by STS measurement methods
Semantic knowledge and corpus used	Semantic knowledge or corpus utilized by STS measurement methods
Strengths and Weakness	STS method capability, determined from aspects such as domain and language independence, the requirement of semantic knowledge, corpus and training data and capability to identify semantic meaning, word order similarity and polysemy
Results	Dataset, experiment setup and result to assess the STS method performance

Fig. 5 Distribution of primary studies per year



based and hybrid-based. This section presents each of them in some detail.

4.1.1 String-based methods

STS methods that fall into this category measure sentence similarity based on the character or string sequence that make the sentences. It does not rely on an external semantic net or corpus to do the similarity calculation. Sentence similarity can be measured by calculating the longest common substring shared by both sentences. Elhadi (2012) introduces a method to calculate text similarity by comparing the longest common sequence between two texts. Sultana and Biskri (2018) propose another method by utilizing the N-grams of characters. N-grams are a subsequence of characters or words that are contained in a sentence or text. First, the method chunks the two sentences being compared into a combination of n-grams of characters with all the possible sizes of n (the maximum result can be achieved by trigram from the experiment). Then it puts the n-grams into a distance matrix for each sentence. A cell in the matrix contains the distance from an n-gram to another n-gram within a sentence. Finally, sentence similarity is measured using the Jaccard (1901) between those two distance matrices. The method is examined in a sentence comparison task following the experimental setup in Takale and Nandgaonkar (2010) and achieved an accuracy of 89.796%. The advantage of this method is that it can be used for any language and domain since it does not rely on semantic ontology or corpus collection. Even though it yields an encouraging result, this method possesses the limitation of not being able to detect passive sentences and semantically similar sentences. Sentence similarity can also be measured by comparing common terms that are shared by both sentences. The Jaccard index is a statistic used in understanding the similarities between sample sets. The measurement emphasizes the similarity between finite sample sets and is formally defined as the size of the intersection divided by the size of the union of the sample sets (Jaccard 1901). A similar approach is used by the Dice coefficient, but it uses a different calculation. The similarity is computed by counting the number of common words, multiplying it by two and dividing by the total number of terms in both sentences (Dice 1945). Salton et al. (1975) introduce a vector space model that can be used for sentence similarity measurement. Sentences are transformed into sentence vectors in the vector space model as illustrated in Fig. 6.

The element of the vector is the terms/words that compose the sentences. Formally, if we want to measure the similarity of sentence D and sentence Q, both sentences can be written as

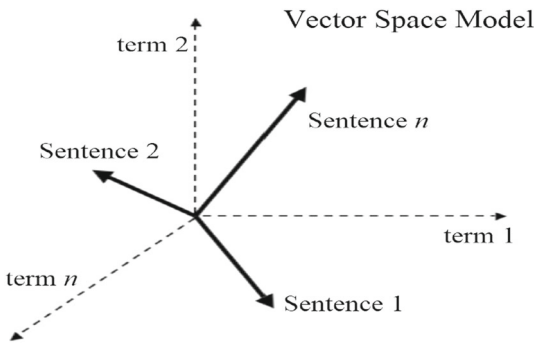


Fig. 6 Vector space model

$$D = (t_0, w_{d_0}; t_1, w_{d_1}; \ldots; t_t, w_{d_t}) \quad (1)$$

$$Q = (q_0, w_{q_0}; q_1, w_{q_1}; \ldots; q_t, w_{q_t}) \quad (2)$$

where t_k represents term and w_{d_k} or w_{q_k} is a weight associated with the term that provides the degree of importance of that term for sentences representation. W_{d_k} is computed using the term frequency inverse document matrix (TF-IDF) scheme from Salton and Buckley (1988).

Salton et al. used cosine vector similarity [Eq. (3))] to measure the sentence similarity.

$$\text{Similarity}\,(Q, D) = \frac{\sum_{k=1}^{t} w_{q_k} \cdot w_{d_k}}{\sqrt{\sum_{k=1}^{t} (w_{q_k})^2 \cdot \sum_{k=1}^{t} (w_{d_k})^2}} \quad (3)$$

4.1.2 Knowledge-based methods

Knowledge-based methods utilize a network of concepts/terms that are semantically interrelated to extract similarity between words before scaling up into sentence level. The semantic network is varied and can be specific to certain domains such as biomedicine and law. However, if it is not available, general-purpose semantic networks such as WordNet can be used. WordNet is a lexical ontology that is similar to a dictionary that contains the concepts or words and its definition (Miller 1995). Every concept or word that has the same meaning is grouped into a synonym set or synset. Each synset is connected in a relationship that forms a semantic network/taxonomy. The relationships can be in the form of "a-part-of," "a-kind-of," "is-the-opposite-of" relationship.

In the literature, we can find numerous formulae to measure the degree of relatedness between the concepts in the semantic network. The path algorithm (Rada et al. 1989) considers the maximum depth of the concepts being compared and the path length in the taxonomy of the semantic network. It uses the following equation to measure the similarity of the concepts.

$$\text{Sim}_{\text{path}}(c_1, c_2) = (2 \times \text{depth}_{\text{max}}) - \text{len}(c_1, c_2) \quad (4)$$

Leacock and Chodorow (LCH) (Leacock and Chodorow 1998) consider the same factor, but it uses the different formula as follows:

$$\text{Similarity}\,(Q,D) = -\log\frac{\text{len}(c_1,c_2)}{2\times \text{depth}_{\max}} \tag{5}$$

On the other hand, Wu and Palmer (1994) include another factor based on the depth of least common subsumer (LCS) in the calculation as follows: LCS itself means the most specific ancestor that the concepts have in common.

$$\text{Sim}_{\text{WP}}(c_1,c_2) = \frac{2\times \text{depth}(\text{LCS}(c_1,c_2))}{\text{depth}(c_1)+\text{depth}(c_2)} \tag{6}$$

Resnik (1995) considers information content in the calculation as defined below.

$$\text{Sim}_{\text{Resnik}}(c_1,c_2) = \text{IC}(\text{LCS}(c_1,c_2)) \tag{7}$$

Information content denotes the specificity of a concept or the probability to find the concept in a corpus. It can be measured by $\text{IC}(C) = -\log(p(c))$, where $p(c) = \text{freq}(c)/N$ and N is the total number of words in the corpus.

Lin (1998) considered the same factor in his calculation, but used a different formula as defined below.

$$\text{IC}(c) = -\log(p(c)) \tag{8}$$

$$\text{Sim}_{\text{Lin}}(c_1,c_2) = \frac{2\times \text{IC}(\text{LCS}(c_1,c_2))}{\text{IC}(c_1)+\text{IC}(c_2)} \tag{9}$$

Jiang and Conrath (1997) also used the following formula.

$$\text{Sim}_{\text{JCN}}(c_1,c_2) = \frac{1}{\text{IC}(c_1)+\text{IC}(c_2)-2\times \text{IC}(\text{LCS}(c_1,c_2))} \tag{10}$$

To scale up to sentence level, we need to devise methods which utilize the above concept similarity measurement.

Liu and Wang (2013) use a vector space model to aggregate word-to-word concept similarity. First, sentence 1 and sentence 2 are transformed into bag-of-word representation. Then, the method forms a joint word set by creating a union of sentence 1 and sentence 2. Semantic vector is formed for each sentence with the joint word set as a vector dimension. Each component of the semantic vector is the maximum similarity value of word pair between every word in the joint word set and every word in a sentence. To measure word pair similarity, they develop their similarity measures based on concept vectors. However, this method could still work using the six-concept similarity that has been discussed previously. After the semantic vector of each sentence is formed, the sentence similarity can be measured by the cosine coefficient of these semantic vectors. They test the method using the experiment setup in Mihalcea et al. (2006). In the paraphrase matching task on Microsoft Research Paraphrase (MSRP) corpus, this method achieves a precision of 0.738 and recall of 0.902. Croft et al. (2013) propose short sentence similarity called lightweight semantic similarity (LSS) which combines the vector space model with path length word-to-word similarity from Rada et al. (1989). The method's step is started by constructing a joint word set from both sentences and using this as a dimension for vector space. The method creates a vector representation of each sentence. Each word in a sentence is compared with each term of joint word set. A word-to-word similarity value between each word with a term in joint word set is summed and used as the value of a vector component related to that term. The step is repeated so that the value of each vector component (term) is obtained. Using a similar process, the method forms vector representation for the other sentence. Overall sentence similarity is calculated by using cosine similarity on sentence vectors. The same experimental setup as in Li et al. (2006) is used to measure the method's performance. Sixty-five noun-pairs from Rubenstein and Goodenough are used, and each word is replaced by its definition from Collin Cobuild dictionary. Then the similarity between the sentence definition of the noun-pair is measured by LSS method and also human judgment. Finally, Pearson's correlation is calculated for both methods. It is reported that the test could achieve a Pearson correlation of 0.807 which is 0.9% lower than the correlation value of the method in Li et al. (2006). However, LSS performs faster due to the omission of word order similarity and corpus statistic.

Li et al. (2012) combine sentence semantic similarity with word order similarity to obtain overall sentence similarity. The method is illustrated in Fig. 7.

The method starts by creating a joint set of words from both sentences. Then the verb and noun vector is created from each sentence by using the word set as a vector dimension. A component of each vector is the similarity value between each word with the joint word set using Lin's word-to-word similarity formula and WordNet. After the noun and verb vector is created, cosine similarity is calculated between vectors of the same type. The result is then combined to obtain semantic similarity.

On the other hand, the word order vector is constructed by assigning a sequence number to each word of both sentences. Word order similarity is obtained by using a certain formula. Finally, the overall sentence similarity is calculated by combining semantic similarity and sentence similarity. To test the method, they apply it to the task to detect sentence similarity using CMU newsgroup dataset. As a result, the method could achieve precision and recall of 0.868 and 0.925, respectively.

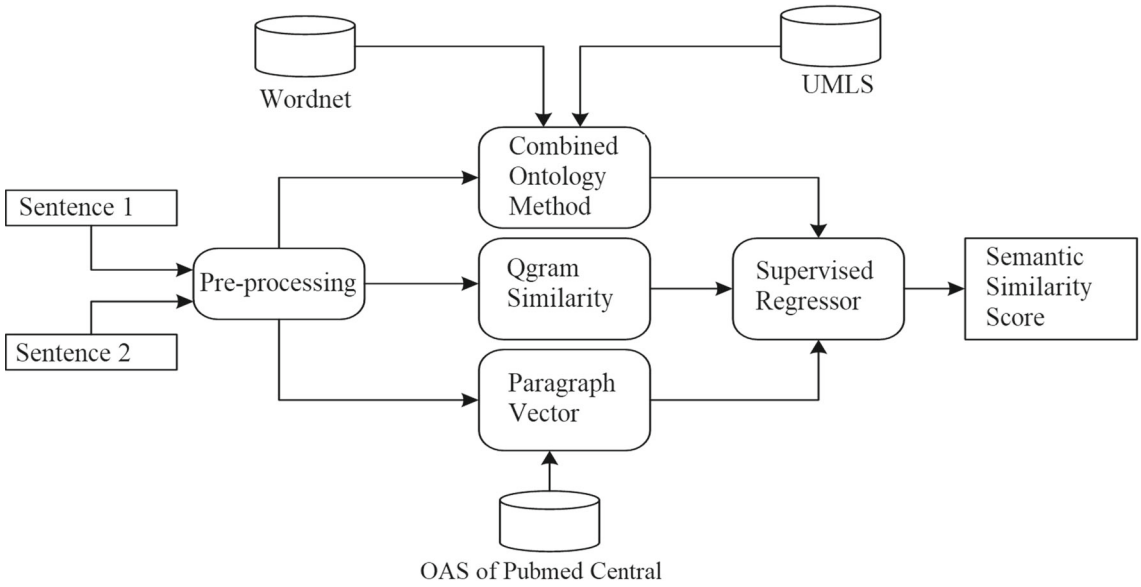


Fig. 7 Sentence similarity measurement

A different approach is taken by Castillo and Cardenas (2010). They tokenize the sentences being compared into two lists of a token. Word by word similarity from both token lists is measured using word similarity from Resnik (1995), Lin (1998), Jiang and Conrath (1997) and Pirró and Seco (2008). Then, the problem of similarity between two lists of words is transformed into bipartite graph matching and solved by using the Kuhn (1955). Finally, sentence similarity is measured by summing optimal assignment in the graph divided by the maximum number of the token between the two lists of the token. The method is tested in the experimental setup of textual entailment recognition. Therefore, the result cannot be compared with the result of Liu and Wang (2013). Nonetheless, the method could achieve maximum accuracy of 56.83%.

Wang and Taylor (2007) use a technique called concept forest as a basis for text similarity. The method starts by extracting keywords from both texts being compared and by stemming the keywords into the base form of the word without inflexion. In each document, each keyword is compared to each other semantically by utilizing WordNet. All terms that can be related in WordNet are grouped forming a treelike hierarchical structure called concept forest. There might be several concept forests within a text. The same process is applied to another text being compared to extract its concept forest. In the process of constructing the concept forest, the methods also consider semantic content rate (SCR) which is the frequency of the term in the text. The text similarity is measured by comparing concept forest from both texts using the Jaccard index. The method is tested in document clustering tasks with a dataset of Reuters-21578. The result can outperform methods that use the vector space model and latent semantic index by the accuracy of 80%.

4.1.3 Corpus-based methods

Corpus-based methods use an external corpus to extract the relation between words or text. Some methods derive the relation between words from a large corpus and then aggregate this relation to measuring similarity in higher extend or sentence level. The other methods can measure text similarity directly without the process of scaling up.

O'Shea et al. (2008) applied the latent semantic analysis (LSA) (Deerwester et al. 1990) to measure text coherence. Initially, it is intended for a large document, but it is also applicable for short text or sentence. LSA assumes that related words will co-occur in the same context/paragraph. LSA derives the relation between words and context from a large collection of a corpus and represents this relation in the form of the word by context matrix. An entry in the matrix means that a word is present in a particular context. The resulting matrix could be in very high dimension which is very computationally expensive. Thus, the matrix dimension needs to be reduced. The method decomposes the matrix using singular values decomposition (SVD) into three others matrices including a diagonal matrix of singular values. This diagonal matrix is truncated by deleting small singular values to reduce its dimension. Then the original word by context matrix is reformed from reduced dimensional space. Each sentence is represented in the form of a vector in the reduced dimensional space to compute sentence similarity. Then the similarity is measured by computing the distance between these vectors

(measured, e.g., with cosine function). LSA can acquire world knowledge that spread in the context. However, this method has several limitations when used to measure short text similarity. First, because of the computational limit of SVD, word by context matrix size is fixed in a certain size. As the effect, there is a possibility that words in the input sentence are not included in LSA dimensional space. Second, the dimension is in a fixed size, so, input sentence will have a very sparse representation. Finally, LSA ignores word order.

Rus et al. (2013) use latent Dirichlet allocation (LDA) (Blei 2003) to measure document/sentence similarity. LDA is a probabilistic approach to model a document into a distribution of topics. This method works by first semi-randomly assigning each word in a document by topics following Dirichlet distribution. In this assignment, each document is represented with topics, and each topic is represented by words. The method will conduct a repeated update of this assignment by considering the proportion of words in a document that are assigned to a topic and proportion of assignments to a topic, overall documents that come from a word. This update will continue it converge to steady-state. As a result, we obtain a document representation in the distribution of topics and topics representation in the distribution of words. Topic distribution of a sentence is compared with the topic distribution of other document using Hellinger distance formula to measure document similarity. Another approach is a word-to-word similarity as a basis and then extend it to the document level. Both documents are represented as bipartite graph matching where words from each document are viewed as vertices and connected by edges as a word-to-word similarity. The document similarity is measured by calculating the optimal assignment between words from both documents represented as a bipartite graph using the Kuhn–Munkres algorithm. Rus et al. test the method in paraphrase detection using MSRP sentence pairs. LDA with bipartite graph matching achieves accuracy and precision of 73.27 and 77.05, respectively.

A different approach is taken by Gabrilovich and Markovitch (2007) by proposing an explicit semantic analysis (ESA) method to measure the relatedness of the text fragment. Figure 8 illustrates how the method works. The method represents text input into an ordered sequence of a weighted vector in a high-dimensional concept extracted from Wikipedia corpus.

Then the semantic relatedness is calculated by comparing vector representation using distance metrics for example cosine coefficient. The authors test the method by using the text pair extracted from ABC news corpus. They calculate the Pearson correlation between the result of their algorithm with human judgment. The method achieved the Pearson correlation score of 0.72 which is better than LSA in the same experimental setup.

Shrestha (2011) proposed a method based on the vector space model (VSM). First, the method builds a term document matrix with the document, like the dimension, constituting the training corpus and a term as a unique term in the training corpus. However, unlike the regular VSM, they reduce the dimension by keeping the dimension with value 1. After the term vectors are obtained, it is used to construct a document vector for the sentence being compared. The term vector is added to the sentence if the term is present. The method also adds a weighting scheme for inverse document frequency to the document vectors. Finally, to measure sentence similarity, cosine similarity is applied to the document vectors. To measure the performance, the authors applied the method to the complete MSRP corpus on the paraphrase detection task. It could achieve an accuracy of 0.683 and precision of 0.703.

Another approach is proposed by Kusner et al. (2015). They leverage word2vec technique by Mikolov et al. (2013) to generate word embedding from Google News corpora. Word embedding means representing words as a dense numerical vector representation. The distance between the word embedded vector is semantically meaningful to a certain extent. The method represents two sentences as normalized bag-of-word vectors to measure the sentence similarity. The distance between the two sentences is measured using the word mover distance (WMD) function. The function calculates the minimum cumulative distance that word in the first sentence needs to travel to match exactly the word in the second sentence. The distance between words is measured using Euclidean distance between the word embedded vectors. The process is shown in Fig. 9. As the final result of WMD computation, the more distance between the two sentences indicates the less similar between two sentences.

This method performs very well in document classification task by outperforming methods using bag of words, TF-IDF (Salton and Buckley 1988), BM25 (Robertson et al. 1995), latent semantic index (Deerwester et al. 1990), latent Dirichlet allocation (Blei 2003), marginalized stacked denoising autoencoder (Chen et al. 2012) and componential counting grid (Perina et al. 2013).

4.1.4 Hybrid methods

Li et al. (2006) proposed a method to calculate sentence similarity by considering semantic and word order information implied in the sentences. To calculate the semantic meaning of the sentences, they combined a knowledge-based and corpus-based method. The proposed method is illustrated in Fig. 10.

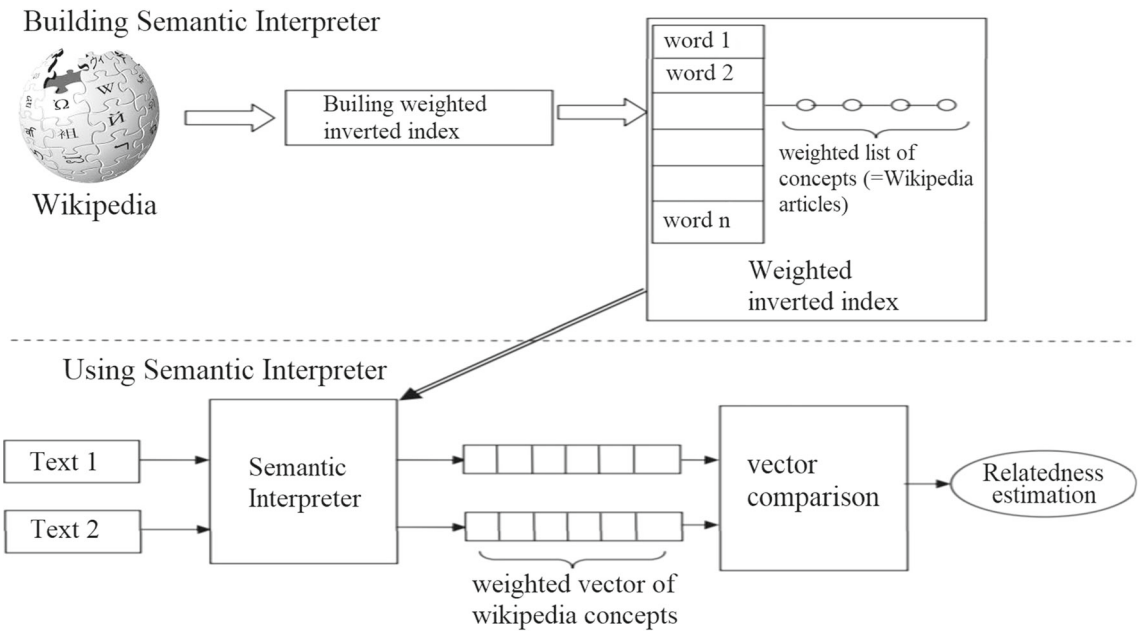


Fig. 8 Principle of explicit semantic analysis

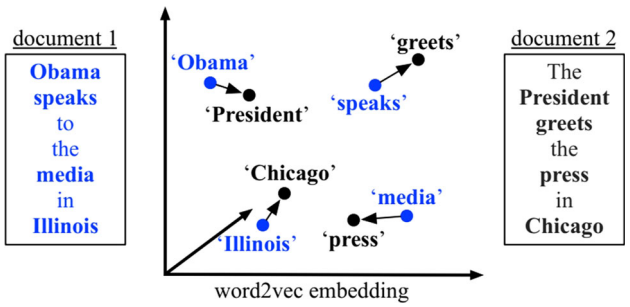
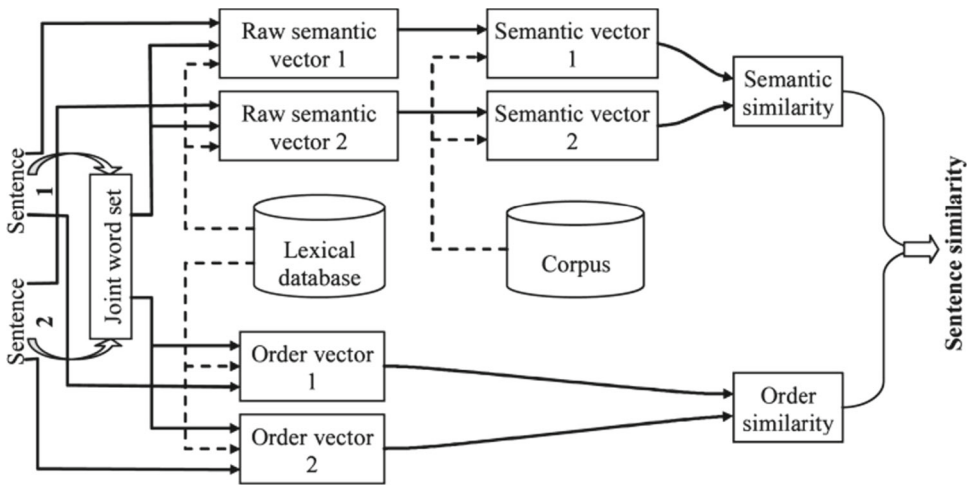


Fig. 9 Sentence similarity using word2vec and WMD

The method combines two input sentences to create a joint word set. Then the input sentences are transformed into a raw semantic vector by using knowledge from a lexical database (WordNet) and joint word set as vocabulary. Each raw semantic vector component will be assigned a value of one if it is present in the joint word set. However, if not, the degree of similarity between words will be calculated by considering the shortest path between the two words and the depth of the subsumer in the WordNet taxonomy.

With a similar mechanism, order vectors are also constructed. Each word in a sentence has a different significance to the meaning of the sentence therefore different weighting must be applied to each word. The method does this by using information content derived from a corpus (Brown corpus). Semantic vectors are formed by combining the raw semantic vector with this information content.

Fig. 10 Sentence similarity methodology



Then semantic and order similarity is calculated for each of the respective vectors. Finally, sentence similarity is calculated by combining semantic and order similarity. The method is tested using Rubenstein and Goodenough word pairs. The same dataset is also presented to human participants and rated by them. Then the rating by the method and the human participant are compared. It was found that the method achieved reasonably good Pearson correlation coefficient of 0.816. Although this method gives a promising result, it has a drawback. In the process where word similarity is measured, word sense disambiguation is not conducted. If the inappropriate sense is used for the words, then the shortest path length between words and the least common subsumer will be falsely calculated, which in turn will lead to incorrect word similarity value.

To overcome this problem, Li et al. (2006) and Pawar and Mago (2018) propose a method that is similar but extends this method by adding word sense disambiguation steps. The method is depicted in Fig. 11. The method starts by partitioning the input sentences into a list of tokens (tokenization). After that, Part-of-speech tagging is applied for each token/word to label them accordingly. A semantic vector is constructed for each sentence which contains the word similarity value assigned to each word for every other word from the second sentence in comparison. The calculation of word similarity is done by utilizing WordNet as a semantic net. This calculation is measured by considering the shortest path length between words and depth of least common subsumer in WordNet as a hierarchy. This process of semantic vector construction also considers information content derived from WordNet as a corpus. The method calculates semantic similarity from these two semantic vectors. As an optional capability, word order vector can be formed to calculate word order similarity. Finally, sentence similarity is measured by combining semantic and word order similarity.

Unlike two previous methods, Islam and Inkpen (2008) use string similarity and corpus-based similarity. For string similarity, they combine three types of modified longest common subsequence and give different weight to each type. They also use second-order co-occurrence PMI (Islam and Inkpen 2006) for corpus-based and word order similarity checking. Similar testing environment as in Li et al. (2006) research is used, and as a result, the method achieved a Pearson correlation coefficient of 0.853 which outperforms Li et al.'s method.

Mihalcea et al. (2006) calculate sentence similarity by aggregating the maximum similarity score between each word of a sentence with each word in the pair's sentence. Then the value is weighted by the inverse document frequency values of each word with the help of British National Corpus. The similarity between words is calculated by combining all six-concept similarity formulae that have been explained in Sect. 4.1.2. They test the method by using a dataset of MSRP. Their method could achieve an accuracy of 0.703.

Vu et al. (2014) use a different approach to measure sentence similarity by combining explicit semantic analysis (ESA) (Gabrilovich and Markovitch 2007) with Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin and Hovy 2003). ROGUE is a lexical similarity measure that is based on n-gram co-occurrence statistic.

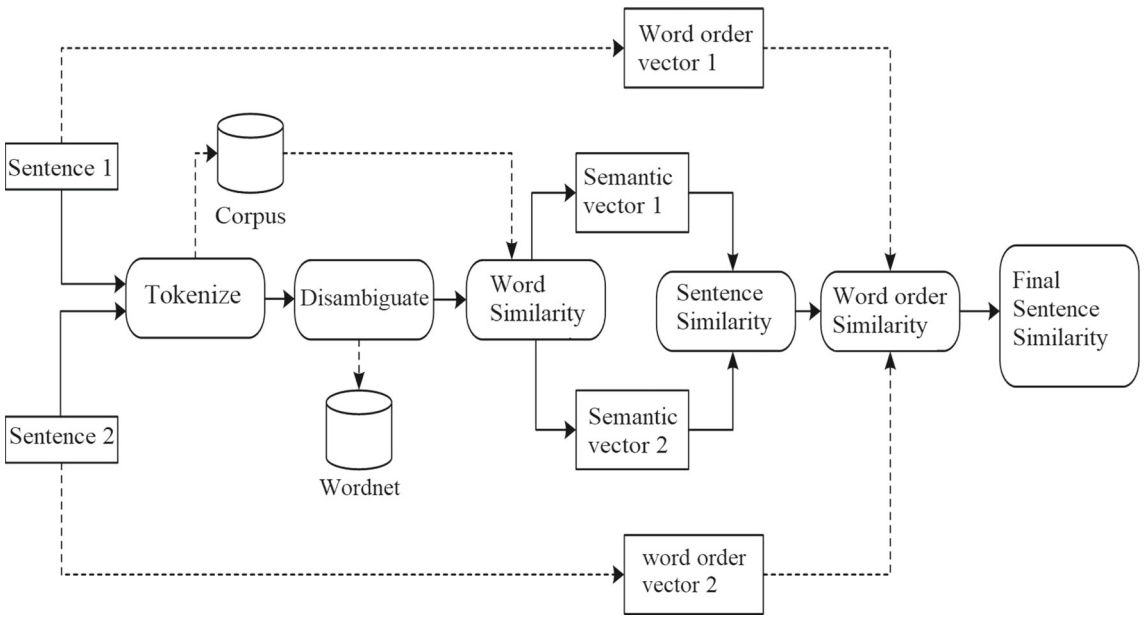


Fig. 11 Sentence similarity methodology

They compute sentence similarity with each method and then calculate the final similarity by using a linear combination and a tuning parameter. They test the method by using their own synthesized dataset from Wikipedia articles. The experimental result shows that it could achieve the highest Person correlation between human-annotated score and the method's score with a value of 0.8265.

In 2012, the Association for Computational Linguistics (ACL) held a Semantic Evaluation (SemEval) workshop focusing on the analysis of diverse semantic phenomena of text. One of the tasks in the workshop is semantic textual similarity, where participants can submit methods to measure the level of equivalence of two sentences semantically (Lin and Hovy 2003). They were interested in the top three best performing methods that have been submitted by the participants. The method from Bär et al. (2012) emerged as the best performing method of the task. It combined numerous measurements including string-based (i.e., greedy string tiling, longest common substring, longest common subsequence, n-grams), knowledge-based [i.e., Resnik (1995) with Mihalcea aggregation function (Mihalcea et al. 2006) to scale up to sentence level], corpus-based (i.e., explicit semantic analysis with Wikipedia and Wiktionary as resource corpus) and two additional text expansion mechanism (i.e., Lexical Substitution System, Statistical Machine Translation). The result of each measurement was used as a feature of a regression model. This model was then applied to detect paraphrases in the union of MSRP and SMT datasets. The result from the method was then compared with the result of manual paraphrase detection by a human participant. The final result showed that the method could achieve an overall Pearson correlation of 0.823.

The second best method is from Šarić et al. (2012). They use a similar approach by combining several sentence similarity measures. Each measure is used as a feature for supervised machine learning models such as support vector regression. The difference is that Šarić et al. (2012) use their sentence similarity measure that comprises of an n-gram overlap, WordNet-augmented overlap, weighted word overlap (with Google Books as a corpus), vector space similarity, shallow named entity recognition and number overlap. The WordNet-augmented overlap is built upon word similarity measurement form Leacock and Chodorow (1998), while the vector space similarity utilizes distributional vector of each word from latent semantic analysis. New York Times corpus and Wikipedia is used as input corpus for LSA. This hybrid method (the simple variant) could achieve a Pearson correlation of 0.813.

Another method proposed by Banea et al. (2012) ranked third best in the same workshop. It combined knowledge-based, corpus-based semantic similarity and bipartite graph matching. The calculation result from each similarity measure is used as features for supervised machine learning technique specifically support vector regression. For the knowledge-based measure, the method uses numerous word-to-word knowledge-based similarity measures that have been discussed in Sect. 4.1.2. To extend it to the sentence level, they make use of an aggregation function from Mihalcea et al. (2006). For corpus-based semantic similarity, the methods exploit latent semantic analysis, explicit semantic analysis and salient semantic analysis. Wikipedia is used as an input corpus to conduct those semantic analyses. The method also incorporates bipartite graph matching to extend beyond the bag-of-word approach. First, each sentence is tokenized into two sets of words as nodes. Each node is connected by an edge which represents word-to-word similarity based on the lexical, semantic and syntactic comparison. A weighing mechanism using a perceptron algorithm is also applied to the graph. After the graph is constructed, the Kuhn (1955) is applied to calculate the optimal alignment. On the same experimental setup as two previous methods, this method could achieve an overall Pearson correlation of 0.7846.

In a specific domain such as biomedicine and law, there is also a need to measure sentence similarity. However, this task has its challenges when the sentence being compared contains many terms which are specific to that domain. Soğancıoğlu et al. (2017) propose a method to measure sentence similarity in the biomedical domain.

Like the previous method we have discussed, this method also combines several sentence similarity metrics and uses the result as a feature for a supervised machine learning method. The method is illustrated in Fig. 12.

After text preprocessing, the method measures the knowledge-based similarity (combined ontology), string similarity (q-gram) and corpus-based (paragraph vector) of each sentence. The result of each measurement is passed to the supervised regression model. The combined ontology measure uses both WordNet, as general-purpose ontology and Unified Medical Language System, as a biomedical ontology to cover biomedical terms that might be excluded in WordNet. On the other hand, for paragraph vector, the method utilizes PubMed that consists of a biomedical corpus to build their vector model. For testing purposes, they synthesize their dataset that consists of biomedical sentence pair. They compare their result with a human-annotated similarity score and calculate the Pearson correlation between the two. The method could achieve a Pearson correlation of 0.836.

4.2 Answer to RQ2

In this section, we explain the strengths and weaknesses of the existing methods. Seven aspects in two groups are used to examine the method's strengths and weaknesses.

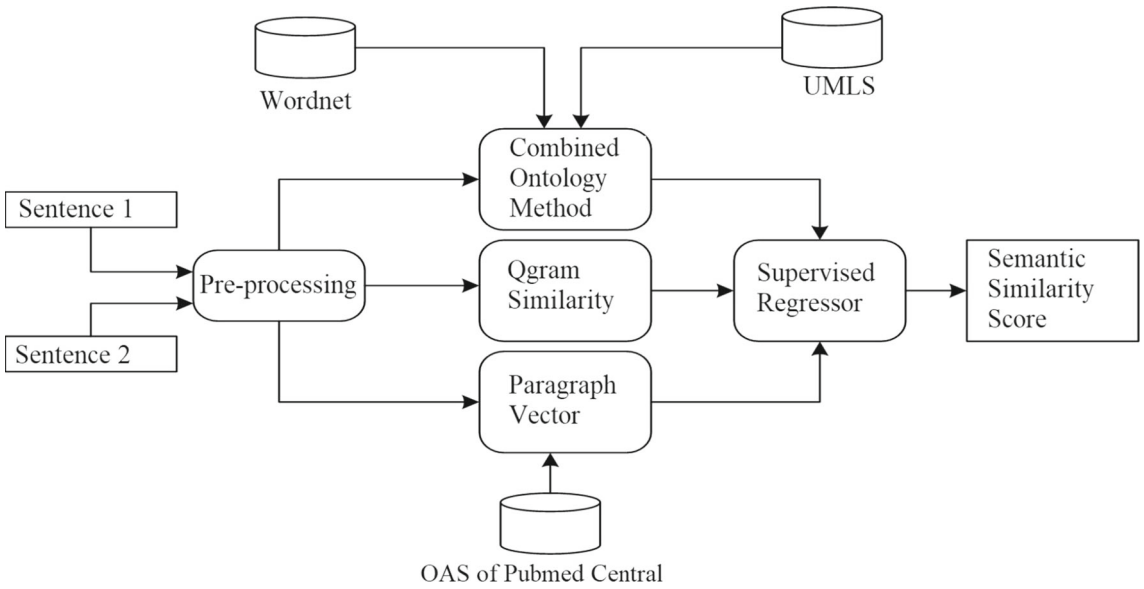


Fig. 12 Sentence similarity in the biomedical domain

The first group is related to the data requirements. The first aspect of this category is the domain and language dependency. This aspect looks at whether the methods can be used for various languages or various specific domains such as biomedicine, chemical or the law. The second aspect is the semantic knowledge requirement. We check whether these methods require semantic knowledge such as from using WordNet, to derive word-to-word similarity, so that the methods could work well. The third aspect is whether the method requires the existence of a corpus to derive relations between words before the relationship can be used for sentence-level similarity measurement, e.g., LSA, LDA, word2vec. The fourth is training data requirements. This aspect will evaluate whether the methods require training and testing data consisting of annotated sentence pairs to function properly.

The second group is related to semantic similarity which consists of three aspects. The first aspect to be examined is semantic meaning. We check the capability of the methods to identify a high degree of similarity from two sentences that have a similar meaning but a different sequence of character or words. For example, "I like that bachelor" and "I like that unmarried man." The second aspect is the word order similarity. This aspect determines the ability of the methods to distinguish the meaning of sentences consisting of the same words but different word order. For example, "Adam hates John" and "John hates Adam." Even though the two sentences comprise the same words, the semantically are different in meaning. The last aspect we review is polysemy, which is a word that has several meanings depending on the context where the word is used. For example, the word "bank" has a particular meaning in an economic context and another meaning in the context of rivers.

We list the findings we obtained from the literature in Table 3. The seven aspects that we have described above are displayed in columns named domain and language independent, requires semantic knowledge, requires corpus, requires training data, semantic meaning, word order and polysemy. The checklist mark in columns 5, 6, 7, 8 means that the aspect is required by the methods, while that in columns 9, 10, 11 means that the capability is presented in the methods.

4.3 Answer to RQ3

To answer this research question, we consider the papers that discuss similarity measurement at the word level in the primary studies, since it could be extended into sentence level by a technique that has been elaborated in Sect. 4.1. As explained in Sect. 4.1, knowledge-based and corpus-based methods require the semantic meaning of words and relations between words obtained from the semantic knowledge and corpus resource to effectively measure sentence similarity. Semantic knowledge takes the form of a network that consists of words or concepts that are interconnected with an explicit relationship such as is-a or part-of. Meanwhile, the corpus is only a collection of documents consisting of words whose relationships are not explicitly known. Certain techniques that combine statistics, probabilities and neural networks such as word2vec, LSA and LDA can be used to derive relationships between words or documents.

Table 3 Strengths and weaknesses of short text similarity measurement methods

Citation	Type	Technique used	Evaluation metric	Domain and language independent	Requires semantic knowledge	Requires corpus	Requires training data	Semantic meaning	Word order	Polysemy	Remark
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Ukkonen (1995)	String	Character matching, suffix trees	–	✔	–	–	–	–	–	–	Cannot identify semantic meaning
Ukkonen (1995)	String	Character matching	–	✔	–	–	–	–	–	–	Cannot identify semantic meaning
Sultana and Biskri (2018)	String	Character matching, N-Gram of characters	–	✔	–	–	–	–	–	–	Cannot identify semantic meaning
Jaccard (1901)	String	Term matching	–	✔	–	–	–	–	–	–	Cannot identify semantic meaning
Dice (1945)	String	Term matching	–	✔	–	–	–	–	–	–	Cannot identify semantic meaning
Salton and Buckley (1988)	String	Term matching, vector space model (VSM), TF-IDF weighting, cosine similarity	–	✔	–	–	–	–	–	–	Cannot identify semantic meaning
Liu and Wang (2013)	Knowledge	JCN word similarity, concept hierarchical model	Paraphrase detection task on MSRP dataset, accuracy = 0.72, precision = 0.738, recall = 0.902, F-measures = 0.8111	–	✔	–	–	✔	–	–	Less accurate than LDA
Croft et al. (2013)	Knowledge	Path word similarity, VSM, cosine similarity	Sentence similarity task on Rubenstein and Goodenough sentence pair dataset, Pearson correlation coefficient = 0.807	–	✔	–	–	✔	–	–	Faster but less accurate than LSA and STASIS
Li et al. (2012)	Knowledge	Information content, cosine similarity on noun and verb vectors	Sentence similarity task on CMU dataset, accuracy = 0.893, precision = 0.868, recall = 0.925, f-measure = 0.896	–	✔	–	–	✔	✔	–	Using their own testing dataset, so the result is incomparable with other research in terms of performance and accuracy

Table 3 (continued)

Citation	Type	Technique used	Evaluation metric	Domain and language independent	Requires semantic knowledge	Requires corpus	Requires training data	Semantic meaning	Word order	Polysemy	Remark
Castillo and Cardenas (2010)	Knowledge	Resnik, Lin, JCN, Pirro word similarity, bipartite graph matching, Hungarian algorithm	Textual entailment recognition task, accuracy = 56.83%	–	✔	–	–	✔	–	–	Text similarity as part of text entailment recognition system
Wang and Taylor (2007)	Knowledge	Concept forest, IDF	Document clustering on Reuters dataset, accuracy = 80%	–	✔	–	–	✔	✔	–	Text similarity as part of Document clustering
Garla and Brandt (2012)	Corpus	Latent semantic analysis (LSA)	Sentence similarity task on Rubenstein and Goodenough sentence pair dataset, Pearson correlation coefficient = 0.838	–	–	✔	–	✔	–	–	More accurate than LSS and STASIS but slower due to its sparse matrix, computationally expensive.
Rus et al. (2013)	Corpus	Latent Dirichlet allocation (LDA)	Paraphrase detection task on MSRP dataset, accuracy = 0.733, precision = 0.771, recall = 0.852, F-measures = 0.81	–	–	✔	–	✔	–	✔	Polysemous words are contained in different topics
Gabrilovich and Markovitch (2007)	Corpus	Explicit semantic analysis (ESA)	Sentence similarity task on ABC news mail dataset, Pearson correlation coefficient = 0.72	–	–	✔	–	✔	–	✔	Using their testing dataset, so the result is incomparable with other research in terms of performance and accuracy
Shrestha (2011)	Corpus	Vector space model, inverse document frequency	Paraphrase detection task on MSRP dataset, accuracy = 0.683, precision = 0.703, recall = 0.917, F-measures = 0.796	–	–	✔	–	✔	–	–	Less accurate than LDA
Kusner et al. (2015)	Corpus	word2vec, earth mover distance	Document classification on 8 dataset, average kNN error 0.42	–	–	✔	–	✔	–	–	Text similarity as part of Documents classification

Table 3 (continued)

Citation	Type	Technique used	Evaluation metric	Domain and language independent	Requires semantic knowledge	Requires corpus	Requires training data	Semantic meaning	Word order	Polysemy	Remark
Li et al. (2006)	Hybrid	Word similarity based on WordNet, information content, word order similarity and cosine similarity.	Sentence similarity task on Rubenstein and Goodenough sentence pair dataset, Pearson correlation coefficient = 0.816	–	✔	✔	–	✔	✔	–	More accurate than LSS
Pawar and Mago (2018)	Hybrid	Words similarity based on WordNet, word sense disambiguation, information content, word order similarity and cosine similarity.	Sentence similarity task on Rubenstein and Goodenough sentence pair dataset, Pearson correlation coefficient = 0.875	–	✔	✔	–	✔	✔	✔	More accurate than STASIS
Islam and Inkpen (2008)	Hybrid	Longest common subsequence, SOC-PMI, word order similarity	Sentence similarity task on Rubenstein and Goodenough sentence pair dataset, Pearson correlation coefficient = 0.853	–	✔	✔	–	✔	✔	–	More accurate than STASIS
Mihalcea et al. (2006)	Hybrid	PMI-IR, LSA, LCH, JCH word similarity, inverse document frequency	Paraphrase detection task on MSRP dataset, accuracy = 0.703, precision = 0.696, recall = 0.977, F-measures = 0.813	–	✔	✔	–	✔	–	–	Less accurate than LDA
Vu et al. (2014)	Hybrid	ROGUE, ESA, N-Grams	Sentence similarity task on Wikipedia dataset, Pearson correlation coefficient = 0.8265	–	–	✔	–	✔	–	✔	Outperform ESA on Person correlation coefficient
Bär et al. (2012)	Hybrid	Longest common subsequence, longest common substring, n-grams, ESA, JCH, Lin, Resnik similarity, linear regression classifier	Paraphrase detection task on MSRP dataset, Pearson correlation coefficient = 0.823	–	✔	✔	✔	✔	✔	–	Best performing method on SemEval 2012

Table 3 (continued)

Citation	Type	Technique used	Evaluation metric	Domain and language independent	Requires semantic knowledge	Requires corpus	Requires training data	Semantic meaning	Word order	Polysemy	Remark
Šarić et al. (2012)	Hybrid	Knowledge- and corpus-based (LSA) word frequencies	Paraphrase detection task on MSRP dataset, Pearson correlation coefficient = 0.813	–		✓	✓	✓	✓	–	Second best performing method on SemEval 2012
Banea et al. (2012)	Hybrid	WordNet-based word similarity, LSA, ESA, SSA, bipartite graph matching, vector regression	Paraphrase detection task on MSRP dataset, Pearson correlation coefficient = 0.785	–	✓	✓	✓	✓	✓	✓	Third best performing method on SemEval 2012
Soğancıoğlu et al. (2017)	Hybrid	Knowledge-based word similarity, Q-gram, word2vec, regression model	Sentence similarity task on Biomedical TAC dataset, Pearson correlation coefficient = 0.836	–	✓	✓	✓	✓	✓	–	Biomedical domain

In the literature, we find numerous semantic knowledge and corpus resources which can be classified into two categories, namely general-purpose and domain-specific resource. The former category means that the resources contain a larger proportion of common words that do not come from specific domains such as dictionaries, newspaper articles and non-scientific books. The latter category is the opposite, where resources mostly contain words or concepts from specific domains such as biomedicine, chemistry and law. General-purpose semantic knowledge corpus are widely used in research to develop general semantic text semantic measurement methods and are not tied to a specific domain such as Croft et al. (2013), Li et al. (2012), Liu and Wang (2013), Kusner et al. (2015), Pawar and Mago (2018), Vu et al. (2014) and Benedetti et al. (2019).

For a semantic knowledge resource, we find a general-purpose resource, namely WordNet, and two domain-specific resources, namely the Unified Medical Language System (UMLS) and the Chemical Entities of Biological Interest (ChEBI). WordNet is a lexical ontology in the English language which was launched by Princeton University in 1985. This resource contains words or concepts along with definitions similar to dictionaries. Each word that has the same meaning is grouped in synsets. WordNet contains 17,000 synsets that are connected to a relationship forming a semantic network (Miller 1995). Various versions of WordNet for other languages are also available such as HowNet for Chinese (Wang et al. 2008; Zhao et al. 2009). In the field of biomedicine, UMLS has been widely used for text semantic similarity that extensively uses many specific words or terms in that field (Soğancıoğlu et al. 2017; Batet et al. 2011; Ben Aouicha and Hadj Taieb 2016; Garla and Brandt 2012; Mabotuwana et al. 2013; Sánchez and Batet 2013). UMLS itself is an ontology which contains concepts in fields that are a combination of several ontologies such as SNOMED-CT, MeSH and Gene Ontology (Bodenreider 2004). UMLS was launched in 1986 and has been hosted by the US National Library of Medicine ever since. In the field of chemistry, ChEBI is often used for text semantic similarity measurement (Ferreira and Couto 2010), because it contains many concepts from domain chemical compounds such as molecular entities, alkanes and alkyl groups (Ferreira and Couto 2010). ChEBI has been developed by the European Bioinformatics Institute since 2002.

For corpus resource, we find research papers that use general-purpose corpora such as Wikipedia (Gabrilovich and Markovitch 2007), British National Corpus (BNC) (Mihalcea et al. 2006), Brown corpus (Li et al. 2006; Pawar and Mago 2018) and a corpus consisting of various news articles from Google News (Kusner et al. 2015), New York Times (Šarić et al. 2012) and Reuters (Benedetti et al.

2019). BNC is a collection of 100 million words gathered from books, newspaper articles, journals and essays and various kinds of writing. BNC was launched in 1994 and maintained by a consortium led by Oxford University (Burnard 2007). Meanwhile, Brown Corpus is a collection containing 1,014,312 American English words from various sources such as novels, journals, articles and from various categories including fiction, religion, government, social and political (Francis and Kucera 1979). This corpus was published by Kucera and Francis of Brown University in 1961. For the specific corpus domain, we found two corpora: PubMed and FindLaw. PubMed is a corpus containing a collection of articles, books and journals of biomedicine compiled by the United States National Library of Medicine since 1966. Each year new documents are added selectively so that by 2018 this corpus has 29.1 million records. The corpus is widely used for text semantic similarity in the field of biomedicine such as in Soğancıoğlu et al. (2017), Chen et al. (2018). On the other hand, FindLaw is a collection that contains 35,000 legal case documents that were crawled from the FindLaw site. This corpus is used for semantic similarity measurement in the field of law (Sugathadasa et al. 2017). In summary, we list all of our findings in Table 4.

5 Discussion

In the result section, we described the methods that have been in the literature. We also categorized these methods into four categories which are string-based, knowledge-based, corpus-based and hybrid-based. Furthermore, we identified the characteristics, strengths and weaknesses of each method as well as the semantic knowledge and corpus resource that can be used. In this section, we will elaborate further on our findings and also the validity of this SLR report.

5.1 Discussion on result

String-based category the advantages and disadvantages of these type of methods are as follows:

First, these methods are relatively straightforward and fast in their implementation because they measure the similarity of sentences simply by comparing sentences directly, without the necessity to refer to semantic knowledge or the relationship between words derived from a corpus. Therefore, it is suitable in a situation where immediate response is an essential factor to consider.

Second, these methods can work at a certain level regardless of domain and language. They are independent of the language used (it does not matter if the sentences compared are composed in English, Chinese, Arabic language) and they can also handle sentences that contain many words in specific fields such as medical, law or other because they only look at the similarity of the sequence of characters or words. Considering these advantages, we view string-based methods as a suitable method for cases that require the immediate response and where semantic knowledge and a corpus are not available.

Third, they can handle a sentence that contains a typographical error. These methods can also detect similarity if there are only a small fraction of letters that are mistyped. These mistyped words may not be present in the semantic knowledge or corpus which can introduce an error in the accuracy of the sentence similarity measurement. However, if the sentence has a different structure in the sequence of characters or words, but has the same meaning, string-based methods cannot comprehend it correctly. This characteristic is a major drawback of string-based category methods that methods in other categories try to solve.

Knowledge-based category uses semantic knowledge to capture semantic meaning from a sentence. General-purpose semantic knowledge such as lexical ontology can be used for this purpose. For more specific domains, coverage of lexical ontology may not be adequate. For example, WordNet is only able to cover 2% of the concepts found in UMLS (Burgun and Bodenreider 2001) which is a biomedicine ontology. The usage of WordNet in conjunction with UMLS can improve the accuracy of semantic text similarity in the field of biomedicine (Soğancıoğlu et al. 2017).

However, semantic knowledge is not always available for all languages or domains. It may exist for languages that are commonly used in the world such as English, Chinese and German, but not all languages have it. Furthermore, building a domain-specific ontology is a hard task and requires expertise in that domain. As an example, to build biomedicine ontology, the developer must know the definition of the term cephem, beta-lactams and the relation that connects the two terms.

Corpus-based category compared to semantic knowledge, the corpus is relatively easy to obtain just by compiling documents, articles, journals or books. For specific corpus domains, it can be done by manually compiling documents in that domain or by crawling a website containing text for particular fields. For instance, to obtain the corpus for the legal domain, one can crawl legal case documents in FindLaw websites (Sugathadasa et al. 2017). The challenge is how to produce the relationships between words or documents from the corpus. The techniques that are commonly used and proved to be empirically reliable include LSA, LDA, word2vec. However, this technique usually requires considerable computational resources because the corpus to be processed is also huge. Therefore, this technique is usually precomputed before the actual

Table 4 Semantic knowledge and corpus resource

Name	Type	Domain	Description	Used in study
WordNet	Semantic knowledge	General	Lexical ontology for the English language	Liu and Wang (2013), Croft et al. (2013), Li et al. (2012), Castillo and Cardenas (2010), Wang and Taylor (2007), Li et al. (2006), Pawar and Mago (2018), Mihalcea et al. (2006), Bär et al. (2012), Šarić et al. (2012), Banea et al. (2012), Sugathadasa et al. (2017)
HowNet	Semantic knowledge	General	Lexical ontology for the Chinese language	Wang et al. (2008), Zhao et al. (2009)
UMLS (SNOMED-CT)	Semantic knowledge	Biomedical	Biomedical ontology	Soğancıoğlu et al. (2017), Garla and Brandt (2012), Sánchez and Batet (2013)
ChEBI	Semantic knowledge	Chemical	Chemical compound ontology	Ferreira and Couto (2010)
Wikipedia	Corpus	General	English articles	Gabrilovich and Markovitch (2007)
New York Times	Corpus	General	News corpus	Šarić et al. (2012)
Google News	Corpus	General	News corpus	Kusner et al. (2015)
Reuters	Corpus	General	News corpus	Benedetti et al. (2019)
British National Corpus	Corpus	General	English document from various source and category	Mihalcea et al. (2006)
Corpus Brown	Corpus	General	English document from various source and category	Pawar and Mago (2018), Li et al. (2006)
PubMed	Corpus	Biomedical	Biomedical documents from journal, books, article	Soğancıoğlu et al. (2017)
FindLaw	Corpus	Law	Law case documents	Sugathadasa et al. (2017)

sentence similarity measurement is done. The other challenge of the corpus-based method based on the bag-of-word models is that the sentence needs to be chunked before further processing and that the word order in the sentence is not maintained. As a result, word similarity cannot be measured.

Hybrid-based category some hybrid methods combine semantic knowledge and corpus so that the accuracy of short text similarity increases such as Pawar and Mago (2018), Vu et al. (2014), Bär et al. (2012) and Soğancıoğlu et al. (2017). However, the processing time will also increase because two sources must be looked up to determine the semantic similarity of the sentence. Therefore, this method is not suitable for time-sensitive cases. Moreover, some techniques such as Banea et al. (2012), Bär et al. (2012) and Šarić et al. (2012) combine several text semantic similarity techniques and use the results as input features for machine learning methods. However, the use of machine learning requires adequate training data to produce accurate predictions. These training data are not always available in the early stages of system implementation.

5.2 Validity threats

Petersen and Gencel (2013) describe four types of validity threats that can occur in software engineering research as well as this SLR. These threats are descriptive validity, theoretical validity, generalizability and interpretative validity. Descriptive validity refers to accuracy based on facts reported by researchers. The researcher should not leave out essential data or include irrelevant or made-up data that can distort the results of the study. Theoretical validity is based on how much theory or theoretical explanation is obtained through research so that it can be trusted and accounted. Generalizability is related to whether research results can be generalized in groups (internal) or across groups (external). Interpretative validity refers to whether the conclusions are taken from the perspective, thought, purpose and experience of the researcher are objective or not (subjective, biased).

We made several efforts to alleviate the several threads described above. In the context of SLR, descriptive validity can be influenced by the inclusion of irrelevant papers or exclusion of highly relevant and high-quality papers. To prevent this, we carefully designed the search string.

Before compiling a search string, we explored the subject of the STS method and then identified the terminology commonly used in this research area to formulate the search term. We also used several synonyms from the search string to expand the scope of the search process. This search string was refined several times before we applied it not only on one database but to four main databases for the computer science field which are IEEE, ACM, Springer and Science Direct. Besides the search strings, search results can also be influenced by the search engine capabilities of the database. Six relevant papers were not found by search engines but through the forward and backwards snowballing technique that we also applied. In terms of the threat to theoretical validity, we compiled a research question that represented the purpose of the research. The answer to the RQs will be the basis to formulate the theory that can describe the phenomenon being observed. To overcome this threat, we use a data extraction form with specific items that are relevant to answer the research questions. This SLR is carried out primarily by the first author. Therefore, there are limitations related to threats that cannot be eliminated. Regarding interpretative validity, there is a potential for subjectivity in the study selection and data extraction, which can introduce a bias to the SLR results. Therefore, we recommend further research to be carried out by involving other researchers in the field. The presence of other researchers is also important in different stages of the SLR, such as refining the search term and discussing the analysis process.

6 Implications

We obtained the following significant implications of the current research: Our study prepares a broad overview of the existing methods and then classifies them into string-based, corpus-based, knowledge-based and hybrid-based methods. It also identifies characteristics, weaknesses and strengths of proposed methods in terms of (i) domain independence, (ii) language independence, (iii) the requirement of semantic knowledge, (iv) corpus and training data, (v) capability to identify semantic meaning, (vi) word order similarity and (vii) polysemy. Furthermore, it also takes into account the comparison between the different text similarity methods and the difference between semantic knowledge sources and corpora for text similarity. Moreover, the current literature review on short text similarity measurement techniques can be served as a guideline and assist novice and new researchers to understand the concept of short text similarity. Also, expert researchers can utilize this review as a benchmark as well as reference to grasp the limitations of current techniques and select the effective STS measurement methods for a particular problem. It can also contribute to the literature by proposing a new research direction for short text similarity measurement methods. For instance, to undertake experiments on larger datasets to compare all existing proposed methods to assess the performance of each method.

7 Conclusion and further work

The short text similarity measurement is an essential component of many applications in natural language processing (NLP) and it has been revealed to be one of the most vigorous areas in NLP. In this paper, we report our research on a systematic literature review on short text similarity measurement methods. Firstly, we identified STS methods and grouped it into four categories based on the characteristics, techniques and external resources (semantic knowledge and corpus) that the methods used. We categorized them into string-based, knowledge-based, corpus-based and hybrid-based. String-based techniques involve comparing character sequences and words in the short text. Knowledge-based techniques use word-level similarity by utilizing semantic knowledge and then scaling up to the sentence level. Meanwhile, the corpus-based techniques derive the relation between words or topics from a corpus by incorporating techniques such as LSA, LDA and word2vec. Then, the resulting model is used for similarity measurements at the sentence level. Several approaches try to combine methods into a hybrid method to achieve improvements inaccuracy. We found six methods for string-based, five methods for knowledge-based, five methods for corpus-based and nine methods for the hybrid technique. Secondly, we examined the characteristics, weaknesses and advantages of these methods using seven aspects which are ((i) domain independence, (ii) language independence, (iii) the requirement of semantic knowledge, (iv) corpus and training data, (v) capability to identify semantic meaning, (vi) word order similarity and (vii) polysemy. Finally, we analyzed the sources of semantic knowledge and corpus resources and categorized it into two types, namely general-purpose and domain-specific. For the semantic knowledge resource, we found four resources, while for the corpus we found eight resources. Our findings from this research contribute by providing other researchers with a better understanding of STS measurement method characteristics and techniques, while also showcasing the advantages and disadvantages of each method. This study can be used as a reference to consider which method is suitable for a particular case.

The findings from the SLR also provide some additional insights; for example, string-based methods are relatively simple and fast compared to other methods, but it is not

appropriate for comprehending the semantic meaning of the text. Meanwhile, knowledge-based methods require semantic knowledge, which is not always available in every language and domain. Semantic knowledge is also expensive to build because expertise is required to develop comprehensive semantic knowledge. However, the methods that fall in this category can comprehend the semantic meaning of the sentence being compared. Moreover, extensions in several methods in this category give the methods the ability to detect word order similarity and polysemy. Corpus-based methods require external corpus resources to make the methods work. To acquire a corpus resource, one can compile the document either manually or by crawling it from websites and then derive the relations between words or topics with techniques such as LSA, LDA or word2vec. However, these techniques require high computational resources, and therefore, the techniques are usually executed in a precomputed way. Then, the resulting model can be used for STS measurement. This method inherently uses bag-of-word models, so it cannot detect word order similarity. We also found that the use of domain-specific semantic knowledge or corpus for the domain-specific STS measurements could improve the accuracy of this method.

The contributions of our study are multifold: (i) broad overview of existing short text similarity measure techniques; (ii) identifying the key challenges and limitations associated with the techniques; (iii) presentation of string-based, corpus-based, knowledge-based and hybrid-based methods; (iv) a reference to consider which method is suitable for a particular case; and (v) the research is applicable to deepen the understanding of short text similarity.

For future work, we aim to perform experiments on a large dataset to compare methods against human ratings of sentence pairs. In other words, we aim to undertake a comparison of the precise measures across one selected dataset to assess the performance comparison.

Compliance with ethical standards

Conflict of interest I hereby and on behalf of the co-authors declare all the authors agreed to submit the article exclusively to this journal and also declare that there is no conflict of interests regarding the publication of this article.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abdi A, Idris N, Alguliyev RM, Aliguliyev RM (2017) Query-based multi-documents summarization using linguistic knowledge and content word expansion. Soft Comput 21:1785–1801

Abualigah LMQ (2019) Feature selection and enhanced krill herd algorithm for text document clustering. Springer, Berlin

Abualigah LM, Khader AT, Al-Betar MA, Alomari OA (2017) Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Syst Appl 84:24–36

Abualigah LM, Khader AT, Hanandeh ES (2018a) Hybrid clustering analysis using improved krill herd algorithm. Appl Intell 48:4047–4071

Abualigah LM, Khader AT, Hanandeh ES (2018b) A novel weighting scheme applied to improve the text document clustering techniques. In: Innovative computing, optimization and its applications. Springer, Berlin, pp 305–320

Agirre E, Diab M, Cer D, Gonzalez-Agirre (2012) A Semeval-2012 task 6: a pilot on semantic textual similarity. In: Proceedings of the first joint conference on lexical and computational semantics-volume 1: proceedings of the main conference and the shared task, and volume 2: proceedings of the sixth international workshop on semantic evaluation. Association for Computational Linguistics, pp 385–393

Alguliyev RM, Aliguliyev RM, Isazade NR, Abdi A, Idris N (2017) A model for text summarization. Int J Intell Inf Technol (IJIIT) 13:67–85

Altszyler E, Sigman M, Slezak DF (2016) Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database CoRR abs/1610.01520

Aouicha MB, Taieb MAH, Hamadou AB (2018) SISR: system for integrating semantic relatedness and similarity measures. Soft Comput 22:1855–1879

Banea C, Hassan S, Mohler M, Mihalcea R (2012) UNT: a supervised synergistic approach to semantic text similarity. In: SemEval '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 635–642

Bär D, Biemann C, Gurevych I, Zesch T (2012) UKP: computing semantic textual similarity by combining multiple content similarity measures. In: SemEval '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 435–440

Batet M, Sánchez D, Valls A (2011) An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 44:118–125. https://doi.org/10.1016/j.jbi.2010.09.002

Ben Aouicha M, Hadj Taieb MA (2016) Computing semantic similarity between biomedical concepts using new information content approach. J Biomed Inform 59:258–275. https://doi.org/10.1016/j.jbi.2015.12.007

Benedetti F, Beneventano D, Bergamaschi S, Simonini G (2019) Computing inter-document similarity with Context Semantic Analysis. Inf Syst 80:136–147. https://doi.org/10.1016/j.is.2018.02.009

Blei DM (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32:267D–270D. https://doi.org/10.1093/nar/gkh061

Budgen D, Brereton P (2006) Performing systematic literature reviews in software engineering. In: ICSE '06. ACM, New York, NY, USA, pp 1051–1052. https://doi.org/10.1145/1134285.1134500

Burgun A, Bodenreider O (2001) Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System, pp 77–82

Burnard L (2007) Reference Guide for the British National Corpus (XML Edition)

Castillo JJ, Cardenas ME (2010) Using sentence semantic similarity based on WordNet in recognizing textual entailment. In: Kuri-Morales A, Simari GR (eds) Advances in artificial intelligence – IBERAMIA 2010, vol 6433. Lecture notes in computer science. Springer, Berlin, pp 366–375. https://doi.org/10.1007/978-3-642-16952-6_37

Chen M, Xu ZE, Weinberger KQ, Sha F (2012) Marginalized Denoising Autoencoders for Domain Adaptation CoRR abs/1206.4683

Chen Q, Kim S, Wilbur WJ, Lu Z (2018) Sentence Similarity Measures Revisited: Ranking Sentences in PubMed Documents. In: BCB '18. ACM, New York, NY, USA, pp 531–532. https://doi.org/10.1145/3233547.3233640

Cordeiro M, Sarmento RP, Brazdil P, Gama J (2018) Evolving networks and social network analysis methods and techniques. Soc Media J Trends Connect Implic 101:8

Croft D, Coupland S, Shell J, Brown S (2013) A fast and efficient semantic short text similarity metric. In: 2013 13th UK workshop on computational intelligence (UKCI), pp 221–227. https://doi.org/10.1109/ukci.2013.6651309

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41:391–407

Díaz I, Ralescu A (2012) Privacy issues in social networks: a brief survey. In: International conference on information processing and management of uncertainty in knowledge-based systems. Springer, Berlin, pp 509–518

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302. https://doi.org/10.2307/1932409

Elavarasi S, Akilandeswari J, Menaga K (2014) A survey on semantic similarity measure. Int J Res Advent Technol 2:389–398

Elhadi MT (2012) Text similarity calculations using text and syntactical structures. In: 2012 7th international conference on computing and convergence technology (ICCCT), December 2012, pp 715–719

Ferreira JD, Couto FM (2010) Semantic similarity for automatic classification of chemical compounds. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1000937

Francis WN, Kucera H (1979) The brown corpus: a standard corpus of present-day edited American English

Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI'07. Morgan Kaufmann Publishers Inc., San Francisco, pp 1606–1611

Garla VN, Brandt C (2012) Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. In: 2012 IEEE second international conference on healthcare informatics, imaging and systems biology, pp 22–22. https://doi.org/10.1109/hisb.2012.12

Gomaa WH, Fahmy AA (2013) A survey of text similarity approaches. Int J Comput Appl 68:13–18

Goth G (2016) Deep or shallow, NLP is breaking out. Commun ACM 59:13–16. https://doi.org/10.1145/2874915

Goyal N, Singh J (2016) A review on resemblance of user profiles in social networks using similarity measures. Int J Comput (IJC) 22:1–8

Islam M, Inkpen D (2006) Second order co-occurrence PMI for determining the semantic similarity of words. In: European Language Resources Association (ELRA), Genoa, Italy

Islam A, Inkpen D (2008) Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans Knowl Discov Data 2:10–25. https://doi.org/10.1145/1376815.1376819

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Société vaudoise des sciences naturelles 37:547–579

Jiang JJ, Conrath DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy

Kaundal A, Kaur A (2017) A review on WordNet and Vector space analysis for short-text semantic similarity. Int J Innov Eng Technol. https://doi.org/10.21172/ijiet.81.018

Kitchenham B (2004) Procedures for performing systematic reviews. Keele 33:1–26

Kuhn HW (1955) The Hungarian method for the assignment problem. Naval Res Logist Q 2:83–97. https://doi.org/10.1002/nav.3800020109

Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ (2015) From word embeddings to document distances. In: ICML'15. JMLR.org, Lille, France, pp 957–966

Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. WordNet: An Electronic Lexical Database 49:265–283

Li Y, McLean D, Bandar ZA, O'Shea JD, Crockett K (2006) Sentence similarity based on semantic nets and corpus statistics. IEEE Trans Knowl Data Eng 18:1138–1150. https://doi.org/10.1109/TKDE.2006.130

Li Y, Li H, Cai Q, Han D (2012) A novel semantic similarity measure within sentences. In: Proceedings of 2012 2nd international conference on computer science and network technology, pp 1176-1179. https://doi.org/10.1109/iccsnt.2012.6526134

Lin D (1998) An information-theoretic definition of similarity. In: Citeseer, pp 296–304

Lin C-Y, Hovy E (2003) Automatic evaluation of summaries using N-gram Co-occurrence statistics. In: NAACL '03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 71–78. https://doi.org/10.3115/1073445.1073465

Liu H, Wang P (2013) Assessing sentence similarity using WordNet based word similarity. JSW 8:1451–1458

Mabotuwana T, Lee MC, Cohen-Solal EV (2013) An ontology-based similarity measure for biomedical data—application to radiology reports. J Biomed Inform 46:857–868. https://doi.org/10.1016/j.jbi.2013.06.013

Majumder G, Pakray P, Gelbukh A, Pinto D (2016) Semantic textual similarity methods, tools, and applications: a survey. Computacion y Sistemas 20:647–665. https://doi.org/10.13053/cys-20-4-2506

Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI'06. AAAI Press, Boston, Massachusetts, pp 775–780

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, vol 26. Curran Associates, Inc, Red Hook, pp 3111–3119

Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38:39–41. https://doi.org/10.1145/219717.219748

O'Shea J, Bandar Z, Crockett K, McLean D (2008) A comparative study of two short text semantic similarity measures. In: Nguyen NT, Jo GS, Howlett RJ, Jain LC (eds) Agent and multi-agent systems: technologies and applications, KES-AMSTA 2008, vol 4953. Lecture notes in computer science. Springer, Berlin, pp 172–181. https://doi.org/10.1007/978-3-540-78582-8_18

Pawar A, Mago V (2018) Calculating the similarity between words and sentences using a lexical database and corpus statistics CoRR abs/1802.05667

Perina A, Jojic N, Bicego M, Truski A (2013) Documents as multiple overlapping windows into grids of counts. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds)

Advances in neural information processing systems, vol 26. Curran Associates, inc, Red Hook, pp 10–18

Petersen K, Gencel C (2013) Worldviews, research methods, and their relationship to validity in empirical software engineering research. In: 2013 joint conference of the 23rd international workshop on software measurement and the 8th international conference on software process and product measurement, pp 81–89. https://doi.org/10.1109/iwsm-mensura.2013.22

Pirró G, Seco N (2008) Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In: Meersman R, Tari Z (eds) On the move to meaningful internet systems: OTM 2008, vol 5332. Lecture notes in computer science. Springer, Berlin, pp 1271–1288. https://doi.org/10.1007/978-3-540-88873-4_25

Pupazan E, Bhulai S (2011) Social networking analytics BMI Paper, VU University Amsterdam, Amsterdam

Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybernet 19:17–30. https://doi.org/10.1109/21.24528

Rawashdeh A, Ralescu AL (2015) Similarity measure for social networks-a brief survey. In: Maics, pp 153–159

Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. Morgan Kaufmann Publishers Inc., Burlington, pp 448–453

Robertson SE, Walker S, Jones S, Hancock-Beaulieu M, Gatford M (1995) Okapi at TREC-3, pp 109–126

Rus V, Niraula N, Banjade R (2013) Similarity measures based on latent Dirichlet allocation. In: Gelbukh A (ed) Computational linguistics and intelligent text processing, CICLing 2013, vol 7816. Lecture notes in computer science. Springer, Berlin, pp 459–470. https://doi.org/10.1007/978-3-642-37247-6_37

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24:513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18:613–620. https://doi.org/10.1145/361219.361220

Sánchez D, Batet M (2013) A semantic similarity method based on information content exploiting multiple ontologies. Expert Syst Appl 40:1393–1399. https://doi.org/10.1016/j.eswa.2012.08.049

Šarić F, Glavaš G, Karan M, Šnajder J, Bašić BD (2012) TakeLab: systems for measuring semantic text similarity. In: SemEval '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 441–448

Shrestha P (2011) Corpus-based methods for short text similarity. Montpellier, France, p 297

Soğancıoğlu G, Öztürk H, Özgür A (2017) BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics 33:i49–i58. https://doi.org/10.1093/bioinformatics/btx238

Stapić Z, López EG, Cabot AG, de Marcos Ortega L, Strahonja V (2012) Performing systematic literature review in software engineering. In: CECIIS 2012-23rd international conference

Sugathadasa K, Ayesha B, Silva Nd, Perera AS, Jayawardana V, Lakmal D, Perera M (2017) Synergistic union of Word2Vec and lexicon for domain specific semantic similarity. In: 2017 IEEE international conference on industrial and information systems (ICIIS), pp 1–6. https://doi.org/10.1109/iciinfs.2017.8300343

Sultana S, Biskri I (2018) Identifying Similar sentences by using n-grams of characters. In: Mouhoub M, Sadaoui S, Ait Mohamed O, Ali M (eds) International conference on industrial, engineering and other applications of applied intelligent systems. Springer International Publishing, Cham, pp 833–843

Tabassum S, Pereira FS, Fernandes S, Gama J (2018) Social network analysis: an overview wiley interdisciplinary reviews. Data Min Knowl Disc 8:e1256

Takale SA, Nandgaonkar SS (2010) Measuring semantic similarity between words using web documents. Int J Adv Comput Sci Appl (IJACSA) 1:10

Ukkonen E (1995) On-line construction of suffix trees. Algorithmica 14:249–260. https://doi.org/10.1007/BF01206331

Vu HH, Villaneau J, Saïd F, Marteau P-F (2014) Sentence similarity by combining explicit semantic analysis and overlapping n-grams. In: Sojka P, Horák A, Kopeček I, Pala K (eds) Text, speech and dialogue, TSD 2014, vol 8655. Lecture notes in computer science. Springer International Publishing, Cham, pp 201–208. https://doi.org/10.1007/978-3-319-10816-2_25

Wang JZ, Taylor W (2007) Concept forest: a new ontology-assisted text document similarity measurement method. In: IEEE/WIC/ACM international conference on web intelligence (WI'07), pp 395–401. https://doi.org/10.1109/wi.2007.11

Wang C, Long L, Li L (2008) HowNet based evaluation for Chinese text summarization. In: 2008 international conference on natural language processing and knowledge engineering, October 2008, pp 1–6. https://doi.org/10.1109/nlpke.2008.4906789

Wohlin C (2014) Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: EASE '14. ACM, New York, NY, USA, pp 38:31–38:10. https://doi.org/10.1145/2601248.2601268

Wu Z, Palmer M (1994) Verbs semantics and lexical selection. Association for Computational Linguistics, Stroudsburg, pp 133–138

Zhao C, Yao X, Sun S (2009) A HowNet-based feature selection method for Chinese text representation. In: Sixth international conference on fuzzy systems and knowledge discovery, pp 26–30. https://doi.org/10.1109/fskd.2009.280