Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus

Issa Atoum*
Faculty of Information Technology
The World Islamic Sciences & Education University
11947 Amman, Jordan

Ahmed Otoom
Independent Researcher,
Amman, Jordan

***Abstract*—Text similarity plays an important role in natural language processing tasks such as answering questions and summarizing text. At present, state-of-the-art text similarity algorithms rely on inefficient word pairings and/or knowledge derived from large corpora such as Wikipedia. This article evaluates previous word similarity measures on benchmark datasets and then uses a hybrid word similarity in a novel text similarity measure (TSM). The proposed TSM is based on information content and WordNet semantic relations. TSM includes exact word match, the length of both sentences in a pair, and the maximum similarity between one word and the compared text. Compared with other well-known measures, results of TSM are surpassing or comparable with the best algorithms in the literature.**

Keywords*—text similarity; distributional similarity; information content; knowledge-based similarity; corpus-based similarity; WordNet

I. INTRODUCTION

Text similarity is a field of research whereby two terms or expressions are assigned a score based on the likeness of their meaning. Short text similarity measures have an important role in many applications such as word sense disambiguation [1], synonymy detection [2], spell checking [3], thesauri generation [4], machine translation [5], information retrieval [6]–[8], and question answering [9].

There are three predominant approaches to compute text similarity. They can be categorized as corpus-based/ distributional semantic models (DSMs), knowledge-based models, and hybrid methods. DSMs are based on the assumption that the meaning of a word can be inferred from its usage (i.e. its distribution in text). It is based on the following hypothesis: linguistic items with similar distributions have similar meanings [10]. Consequently, these models derive vector-based representations of the meaning of a word co-occurrence in a corpus. The vector-based representation is most often built from large text collections [5]. In this category, the latent Dirichlet allocation (LDA) assumes that each document is based on a mixture of topics, whereas a topic probabilistically generates various words [6], [11]–[13] . In the same category, the latent semantic analysis (LSA) is based on that the words that share similar meaning tend to occur in similar texts [6], [9], [14], [15].

TABLE I. TEXT SIMILARITY EXAMPLE

#	Sentence pairs	Human Score	LSA[a]	Li [16]	Mohler [17]
1	*The cord is strong, thick string.* *A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.*	0.01	0.19	0.33	0.45
59	*A cock is an adult male chicken.* *A rooster is an adult male chicken.*	0.86	1.00	0.83	1.00

[a.] Using TASA Space

The knowledge-based methods usually employ taxonomic information (e.g. WordNet) to estimate semantic similarity [18][19]. Sentence knowledge-based methods use semantic dictionary information such word relationships [19]–[21], information content [22], [23], parts of speech [18], [24], word senses [25], [26], and gloss definitions from a corpus [27], [28] to get the overall semantic score. These methods suffer from the limited number of general dictionary words, which are commonly used in general English literatures and may not suit specific domains.

Hybrid methods integrate various knowledge-based and/or corpus-based methods. They generally perform better [29]. In recent years, much of the work on lexical semantics has focused on distributional vector representation models [30], [31].

We have identified three cases where knowledge-based, corpus-based or *traditional* hybrid methods perform poorly. We illustrate these cases by examples. Table I shows two examples of two sentence-pairs taken from STS-65 benchmark dataset [16] that were compared using: LSA [15] (i.e. corpus-based method), [16] (i.e. knowledge-based method) and [17] (i.e. hybrid method).

The first case is as follows: methods that depend on a large corpus tend to overestimate relatively unrelated sentences or relatively related sentences (e.g., LSA). For the first sentence-pair, we obtained a similarity score of 0.19 (relatively high) for LSA measure, whereas the reported human similarity score

mean is 0.01. The LSA method depends on words' frequencies that tend to be relatively high in a large corpus (e.g., TASA). The second case is as follows: knowledge-based methods have the same drawback as the previously discussed method (LSA). The method of [16] depends on WordNet semantic relations (i.e. path and depth). This method can distinguish between general and specific concepts using WordNet but does not have information about words' distributions (or context). The third case is as follows: *traditional* hybrid methods that combine multiple measures over an average function generally perform poorly [17]. From [17], we determined that each sub-similarity method diverges in score compared to the overall similarity score. Each of the eight different measures has its strengths and weakness and thus will not get an acceptable semantic score in all cases. In many cases, one measure will have high similarity (e.g., >0.5 for LSA) and low similarity (e.g., <0.1 for path measure) over STS-65 dataset. In the second sentence-pair the same finding could be deduced. We deduced that the LSA and [17] measures overestimate the similarity score of the compared sentence-pair. Therefore, a similarity measure that use minimum data resources and get acceptable score is looked for.

Our work presents a hybrid-based text similarity measure that utilizes WordNet [32] information and a corpus[33]. The WordNet is a man-made ontology that shows promising results in the text similarity domain. The proposed method uses a small size word corpus, thereby eliminating the processing of large corpora. Using the weighted word similarity [34], a new text similarity measure is proposed. The proposed measure compares short text to long text and finds the maximum word similarity and the total exact matching words. The final similarity is calculated using the total similarity of the comparable words weighted by the text length in words.

First, the related works are summarized. Next, the proposed approach is presented and explained. Then, the proposed method is evaluated; finally, the article is concluded.

II. Related Work

Sentence similarity methods (also called short text similarity) are used to measure word similarities in a sentence to reflect the overall semantic of the compared sentences. In general, sentence similarities can be categorized as corpus-based, knowledge-based, and hybrid methods.

A. Corpus-based Methods

Corpus models learn word co-occurrence from large corpora to predict the similarity of comparing text. Many models use information from internet sources such as: Wikipedia [35], Google Tri-grams [5], [36], and Search Engine documents [37].These models can be categorized as DSMs and distributed vector representation models.

DSMs derive vector-based representations of the semantic meaning of patterns of word co-occurrence in corpora. In this category, LSA is based on that the frequency of words in certain contexts that could determine the semantic similarity of words to each other. That is, words that are similar tend to occur in similar texts [6], [9], [14], [15]. In latent Dirichlet allocation (LDA) each document is based on a mixture of topics, whereas a topic probabilistically generates various words [6], [11]–[13]. The idea of the vector space model (VSM) [38] is to represent each document in a collection as a point in a space (a vector in a vector space). Points that are close together in the space are semantically similar, whereas points that are far apart are semantically different. The construction of a suitable VSM for a particular task is highly parameterized, and there appears to be little consensus over which parameter settings to use [39]. Moreover, many of these models are based on large corpora. The global vector model (GloVe) is an unsupervised learning model for word representation [40], which is trained on the non-zero elements in a global word–word co-occurrence matrix. The distributional model [41] combines visual features with textual ones, resulting in a performance increase. The explicit semantic analysis (ESA) represents the meaning of any text as a weighted vector of Wikipedia-based concepts [42]. Furthermore, the distributional method of LSA [43] is enhanced with WordNet semantic relations.

Distributed vector representation of words can capture syntactic and semantic regularities in language and help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words. The unified architecture of NLP [44] learns features relevant to the tasks at hand given very limited prior knowledge. This is achieved by training a deep neural network, building upon work by [30], [45] . Their models [44], [46] learn word representations in a binary classification task (related word to its context or not) . They use the learned word representations to initialize the neural network models for other NLP tasks that also have word representation layers. One of the recent works on distributed representations is the work of [31] wherein they used probabilistic feed-forward neural network language model to estimate word representations in vector space. Align, disambiguate, and walk (ADW) model is a graph-based approach that has two steps; word transformation to the word senses (i.e. one of the meanings of a word) and disambiguation by taking context of compared words [47]. Based on WordNet, [48] exploit semantic representations of sentences using extracted features from a logic prover.

B. Knowledge-based Methods

Sentence knowledge-based methods use semantic dictionary information such word relationships [19]–[21], information content [22], [23], word senses [25], [26], and gloss definitions from a corpus [27], [28] to get word semantics. Based on human comprehension of sentence meaning, [49] proposed to measure the sentence similarity from three aspects that people identify in a sentence. People obtain information from a sentence on three aspects, or some of them: *objects* the sentence describes, *properties* of these *objects* and *behaviors* of these objects. Consequently, they propose three similarities: objects-specified similarity, objects-property similarity, objects-behavior similarity, and overall similarity.

Some similarity models [19] measurs the semantic relatedness between texts based on their implicit semantic links extracted from a thesaurus. Other models [25] measures sentence similarity based on word sense disambiguation and

WordNet synonym expansion. They build word sense disambiguation by using gloss interactions and expand it by synonyms. Then, the sentence is similarly calculated using cosine vectors. The reference [50] proposed a sentence similarity that used weighted word noun and verb vectors along with the order of words in a text.

In general, the knowledge-based approach is limited to the use of human-crafted dictionaries. Because of this, not all words are available in the dictionary and even though some word exists, they do not have full semantics.

C. Hybrid-based Methods

Hybrid-based methods are combinations of the previously mentioned methods. The reference [16] proposed a sentence similarity based on a non-linear function of WordNet path and depth, associated with information content form Brown Corpus, and sentence word orders. The reference [7] proposed a weighted similarity vector based on shortest path and term frequency to replace [16] semantic vector. They applied the similarity measure on photographic description data. The weighted textual matrix factorization (WTMF) model [11] is built on WordNet, Wiktionary, and Brown corpus. The reference [18] generated a semantic vector space using part of speech and WordNet. The reference [51] proposed a sentence similarity measure for paraphrase recognition and text entailment based on WordNet for existing words and an edit distance for proper nouns. The reference [24] proposed sentence similarity based on WordNet Information Content and part of speech tree kernels.

The reference [29] proposed a three-layer sentence measure: lexical layer, syntactic layer, and semantic layer. The overall sentence measure depends on the number of tokens, RDF triples that entail the semantic layer. In the same area, [52] combined the words meanings and phrase context in a sentence measure. The meaning words are implied by extracting words' lemma from a dictionary, whereas phrase context usage was extracted using a huge para-phrase alignment database [53].

Many hybrid methods are supervised models. They predict test sentence prevalence to training data. UNT model [54] uses regression machine learning based on hybrid text similarity methods of [17], [55], [56]. UKP system, which performed the best in the Semantic Textual Similarity (STS) task at SemEval-2012, uses the log-linear regression model to combine multiple text similarity measures of varying complexity. The reference [57] proposed the yiGou model. They used the support vector machine model with literal similarity, shallow syntactic similarity, WordNet-based similarity, and latent semantic similarity to predict the semantic similarity score of two short texts. The Takelab model [58] uses support vector regression model with multiple features measuring word-overlap similarity and syntax similarity to predict human sentence similarity. Each sentence is represented as a vector in the LSA model based on word vectors. Hybrid approaches show promising results on benchmark datasets.

III. Proposed Method

We highlighted the imperfections of word similarity measures [34] that are either distance (knowledge)-based [16] or information content (IC)-based [22]. Distance-based methods suffer from the problem of having the same similarity value for words that share the same path or depth in a taxonomy such as WordNet. In contrast, the problem with IC measures is its limitation of available words in a corpus or getting the same similarity when the compared words has the same LCS ratio. We borrow the word similarity of [34] as shown in (1). Furthermore, we modified the word similarity factor of [34] as shown in (2).

$$Sim_{JDIC}(w_{i,}w_j) = \psi \cdot SimA \cdot SimB, \quad (1)$$

where $SimA = log_2(Sim_{Li}(w_{i,}w_j) + 1)$, and

$SimB = log_2(Sim_{Lin}(w_{i,}w_j) + 1)$,

where $w_{i,}w_j$ are compared words, $\psi \in [0,1]$ is a weighting factor that combines the IC of the pairs, and Sim_{Li}, Sim_{Lin} is the word similarity as in Li, Lin.

$$\psi = 1 - e^{-(log_2(IC(w_i)*IC(w_j)+1))}, \quad (2)$$

where $w_{i,}w_j$ are compared words, $\psi \in [0,1]$ is a weighting factor that combines the IC of the pairs, and $Sim_{Li}$and Sim_{Lin} is the word similarity as in Li [16] and Lin [22] respectively.

This article proposes a novel text similarity measure (TSM) that facilitates word similarity in (1). The TSM finds the maximum word similarity and the total exact matching words between compared sentences. Then, the total similarities of compared words are summed up and weighted by sentences' length and a logarithmic function.

The proposed maximum similarity of a word w and a text R is shown in (3).

$$\mathrm{Sim}(w, \mathrm{R}) = arg\ max_{1 \le i \le |R|}\ Sim_{JDIC}(w, R_i), \quad (3)$$

where R_i is the word i in text R and Sim_{JDIC} as defined in (1).

From [1], [33], [58], we inferred that compared text lengths and exact matches words have a direct effect on the final similarity score. The longer the compared text, the higher the chances of getting similar words.

The proposed TSM between two text fragments T, R is shown in (4).

$$\frac{\sum_{i=1,}^{|\mathrm{T}|} \mathrm{Sim}(w_i, \mathrm{R}) . \mathrm{Log}(2.\delta + 2.Max(|\mathrm{T}|, |R|))}{|\mathrm{T}| + |R|} \quad (4)$$

where, δ represents the exact word match between compared sentences. The Max function computes the maximum length between the compared sentences. The Sim function, as defined in (3), stands for the maximum similarity between a word and compared text fragment.

The application of (3) and (4) can be shown by the following sentence-pair taken from STS-65 dataset [16]:

S1: A boy is a child who will grow up to be a man.

S2: A rooster is an adult male chicken.

When we compare the two sentences using (3), the

maximum similar word-pairs from the sentence (S1) to the sentence (S2) are as follows: the word *boy* to the word *male* (0.282), the word *child* to the word *male* (0.153), and the word *man* to the word *adult* (0.786). The length of both sentences is 4 after stemming and removing stop words. Thus, applying (4) we got the similarity of 0.152. Compared to the reported human mean score (0.11), the proposed method got an acceptable similarity score.

IV. Evaluation and Experimental Results

The evaluation of word and sentence measures are as follows.

A. Word Similarities

We evaluated the word similarity [34] on a relatiely small benchmark datasets [60], [61]. Below, we extend the comparison to larger benchmark datasets: WordSim (WS)-353 [62], MEN dataset [63], and SimLex-999 [64]. The WordSimilarity-353 *test collection* contains two sets of English word pairs along with human-assigned similarity judgements. All the subjects in both experiments possessed near-native command of English. Their instructions were to estimate the relatedness of the words in pairs on a scale from 0 (totally unrelated words) to 10 (very much related or identical words). The MEN *test collection* contains two sets of English word pairs (one for training and one for testing) together with human-assigned similarity judgments, obtained by crowdsourcing using Amazon Mechanical Turk via the CrowdFlower interface. The MEN data set consists of 3,000 word pairs, randomly selected on scales 1 (lowest) to 7 (highest) similarity. The SimLex-999 comprises 666 Noun-Noun pairs, 222 Verb-Verb pairs and 111 Adjective-Adjective pairs. SimLex-999 is challenging dataset for computational models to replicate. In order to perform well, they must learn to capture similarity independently of relatedness/association.

The Spearman correlation between different methods is shown Table II. The LSA, [65], [44], and VSM correlation were taken from [64]. We used Brown corpus and WordNet 3.0 for the JDIC measure, Li, and Lin measures. According to the results, both Li and Lin methods perform poorly which links to our initial hypothesis that a (corpus-based or knowledge-based) similarity method often does not perform well. In general, word similarity measures vary from one method to another depending on method features. Some methods use all word tags, while others use nouns only. Moreover, some methods support disambiguation or use additional domain information. The Spearman correlation of the JDIC method got the highest correlation for the SimLex-999 dataset. The JDIC approach looks for similar words and the SimLex-999 dataset is composed of similar words rather than related words. The WS-353 [62] list contains pairs that are associated but not similar in the semantic sense, for example: *liquid – water*. The list also contains many culturally biased pairs, for example: *Arafat – terror* [4]. Nevertheless, on average the borrowed method (JDIC) method achieved acceptable results compared with results of the state-of-the-art methods as shown in figure I. However, without a real system the comparison remains questionable.

We showed that the semantic similarity measures [66] could play a major role in software quality detection. Therefore, we will confirm this finding in the next section by using JDIC in a new text similarity measure.

TABLE II. Spearman Correlation of Word Similarity Measures Over Different Methods

Method/Dataset	MEN	SimLex-999	WS-353
Lin [22]	0.25	0.27	0.27
Li [16]	0.27	0.28	0.24
Huang [65]	0.30	0.10	0.62
VSM [39]	0.43	0.20	0.40
LSA[67]	0.48	0.23	0.40
Collobert [44]	0.57	0.27	0.49
Mikolov [68]	0.43	0.28	0.65
Islam [36]	0.72	0.33	0.62
JDIC [34]	**0.56**	**0.53**	**0.61**
Pennington [40]	0.66	0.40	0.67

B. Text Similarities

Table III shows the Pearson correlation of a list of text similarity methods over the benchmark dataset of Sem-Eval 2012 [69]. The dataset comprises pairs of sentences drawn from publicly available datasets that have been manually tagged with a number from 0 to 5:

- MSR-Paraphrase, Microsoft Research Paraphrase Corpus, 750 pairs of sentences.

http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/

- MSR-Video, Microsoft Research Video Description Corpus, 750 pairs of sentences.

http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/

- SMTeuroparl: WMT2008 development dataset (Europarl section), 734 pairs of sentences.

http://www.statmt.org/wmt08/shared-evaluation-task.html

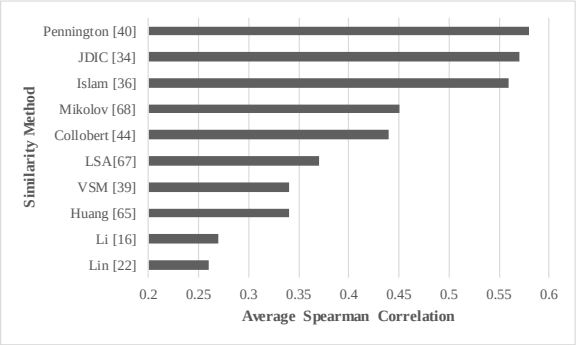

Fig. 1. Average spearman correlations over word similarity methods

Table III shows that the proposed method (TSM) was significant ($p < 0.01$) over all datasets except for a few. We implemented Li sentence measure while the Lin method was implemented based on the sentence measure proposed by [55]. The Google Tri-grams method [36] does not perform as well as for text similarity compared with its performance for word similarity measure (Table II). This finding shows that the word similarity on its own does not always lead to a good text similarity measure. It also supports our hypothesis that measures that use large collection of data could overestimate

unrelated sentences (showed in Table I). The low performance of Li measure was related to its inability to capture relatedness in compared sentences. Preliminary research showed that path and depth alone (Li measure) cannot give better semantic relatedness. Contrariwise, the Lin text similarity method shows an average Pearson correlation of 0.51; thus, the information content gained better similarity scores. Further comparisons on the STS-65 dataset can be found at the work of [70].

TABLE III. PEARSON CORRELATIONS OF SEVERAL METHODS OVER THE SEM-EVAL 2012 DATA SETS

#	Method/Dataset	MSRvid	MSRpar	SMTeuroparl	Sur.OnWN	Sur.SMTnews
1	G. Tri-grams [36]	0.47	0.32	0.41	0.65	0.38
2	Li [16]	0.42	0.42	0.54	0.58	0.34
3	LDA	0.77	0.27	0.45	0.62	0.37
4	ESA [42]	0.75	0.43	0.38	0.62	0.33
5	Lin [22]	0.56	0.55	0.57	0.58	0.27
6	LSA [43]	0.66	0.36	0.57	0.66	0.39
7	ADW [47]	0.80	0.51	0.50	0.54	0.45
8	UNT [54]	0.88	0.54	0.42	0.67	0.40
9	WTMF [11]	0.84	0.41	0.51	0.73	0.44
10	**TSM**	**0.83**	**0.58**	**0.45**	**0.66**	**0.42**
11	yiGou [57]	0.84	0.51	0.48	0.67	0.48
12	Takelab [58]	0.86	0.70	0.36	0.70	0.47
13	UKP [71]	0.87	0.68	0.53	0.66	0.49

We noted high performance of TSM (Pearson 0.66) on the dataset of OnWN because WordNet is one resource of TSM. The TSM performs better than methods (1–9) because each of them is considered to use one technique (knowledge-based or corpus-based) compared to TSM (hybrid). The application of TSM on the two sentence-pairs in Table I got the scores (0.002,0.80), thus our proposed TSM does not adhere to discussed drawbacks of knowledge-based and corpus-based measures. We found that the major performance of TSM was because of the proposed text similarity measure and the borrowed word similarity measure.

However, our method has some limitations. Compared with methods (11–13), it has lower performance. The main reason is that the top scoring methods tend to use most of the available resources and tools. For example, the yiGou 2015 adds the LSA features along with WordNet Similarity features. The TakeLab method uses multiple features that include syntax similarity which is not part of TSM. The UKP method uses a combination of approximately 20 features. These features include n-grams, ESA vector comparisons, and word similarity based on lexical-semantic resources. Furthermore, the TSM could not disambiguate words in different contexts. Therefore, we deduce that our method performance is accepted as it utilizes limited data resources .On average (figure II) our proposed TSM method got an acceptable Pearson correlation. The proposed method may be used in applications that do not require high accuracy such as in search engines or on systems that has low resources such as mobile applications.

V. CONCLUSION

This article presented a new text similarity measure based on previously proposed joint distance and information content word similarity measure, and the information content of compared words. The proposed text similarity is weighted based on comparable text length and the total exact word matches. The similarity measure outperforms much of the compared similarity measures and is significant at the 0.05 level. The reason behind the high achievement of our method is due to the employment of additional information (corpus and information content) and the effectiveness of the borrowed word similarity measure. Although the proposed method has low performance compared to some compared models, it has less machinery and uses low information resources. In future, we plan to apply the proposed method on a real application of software quality.

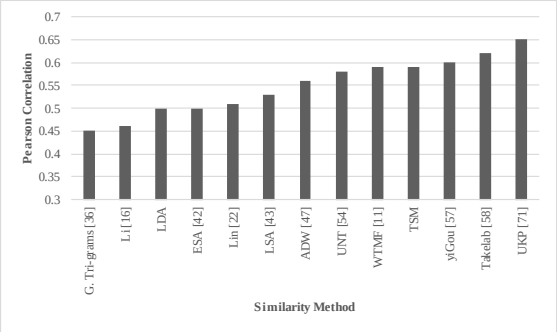

Fig. 2. Average pearson correlations over text similarity methods

ACKNOWLEDGMENT

I would like to thank Prof. Dr. Narayanan Kulathuramaiyer for his valuable feedback and comments. I would like also to thank the anonymous reviewers for their comments.

REFERENCES

[1] C. Ho, M. A. A. Murad, R. A. Kadir, and S. C. Doraisamy, "Word sense disambiguation-based sentence similarity," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, no. August, pp. 418–426.

[2] P. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," in Proceedings of the 12th European Conference on Machine Learning, 2001, pp. 491–502.

[3] A. Islam and D. Inkpen, "Real-word Spelling Correction Using Google Web IT 3-grams," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3, 2009, pp. 1241–1249.

[4] M. Jarmasz and S. Szpakowicz, "Roget's Thesaurus and Semantic Similarity," Recent Adv. Nat. Lang. Process. III Sel. Pap. from RANLP 2003, vol. 111, 2004.

[5] A. Islam and D. Inkpen, "Unsupervised Near-Synonym Choice using the Google Web 1T," ACM Trans. Knowl. Discov. Data, vol. V, no. June, pp. 1–19, 2012.

[6] B. Chen, "Latent topic modelling of word co-occurence information for spoken document retrieval," in IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2009, 2009, no. 2, pp. 3961–3964.

[7] D. Croft, S. Coupland, J. Shell, and S. Brown, "A fast and efficient semantic short text similarity metric," in Computational Intelligence (UKCI), 2013 13th UK Workshop on, 2013, pp. 221–227.

[8] S. Memar, L. S. Affendey, N. Mustapha, S. C. Doraisamy, and M. Ektefa, "An integrated semantic-based approach in concept based video retrieval," Multimed. Tools Appl., vol. 64, no. 1, pp. 77–95, Aug. 2011.

[9] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "A Comparative Study of Two Short Text Semantic Similarity Measures," in Agent and Multi-Agent Systems: Technologies and Applications, vol. 4953, N. Nguyen, G. Jo, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2008, pp. 172–181.

[10] Z. S. Harris, "Distributional structure," Word, vol. 10, no. 2–3, pp. 146–162, 1954.

[11] W. Guo and M. Diab, "A Simple Unsupervised Latent Semantics Based Approach for Sentence 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, 2012, pp. 586–590.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[13] J. Xu, P. Liu, G. Wu, Z. Sun, B. Xu, and H. Hao, "A Fast Matching Method Based on Semantic Similarity for Short Texts," in Natural Language Processing and Chinese Computing, Y. Zhou, Guodong and Li, Juanzi and Zhao, Dongyan and Feng, Ed. Chongqing, China: Springer Berlin Heidelberg, 2013, pp. 299–309.

[14] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse Process., vol. 25, no. 2–3, pp. 259–284, 1998.

[15] S. Deerwester, S. S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," J. Am. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, Sep. 1990.

[16] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," IEEE Trans. Knowl. Data Eng., vol. 18, no. 8, pp. 1138–1150, Aug. 2006.

[17] M. Mohler and R. Mihalcea, "Text-to-text Semantic Similarity for Automatic Short Answer Grading," in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 567–575.

[18] M. C. Lee, "A novel sentence similarity measure for semantic-based expert systems," Expert Syst. Appl., vol. 38, no. 5, pp. 6392–6399, 2011.

[19] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, "Text relatedness based on a word thesaurus," J. Artif. Intell. Res., vol. 37, pp. 1–38, 2010.

[20] N. Seco, T. Veale, and J. Hayes, "An Intrinsic Information Content Metric for Semantic Similarity in WordNet," in Proceedings of the 16th European Conference on Artificial Intelligence, 2004, no. Ic, pp. 1–5.

[21] G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," in WordNet: An electronic lexical database, vol. 305, C. Fellbaum, Ed. Cambridge, MA: The MIT Press, 1998, pp. 305--332.

[22] D. Lin, "An information-theoretic definition of similarity," in Proceedings of the 15th international conference on Machine Learning, 1998, vol. 1, pp. 296–304.

[23] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI'95), 1995, vol. 1, pp. 448–453.

[24] Y. Tian, H. Li, Q. Cai, and S. Zhao, "Measuring the similarity of short texts by word similarity and tree kernels," in IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT), 2010, pp. 363–366.

[25] K. Abdalgader and A. Skabar, "Short-text similarity measurement using word sense disambiguation and synonym expansion," in AI 2010: Advances in Artificial Intelligence, vol. 6464, J. Li, Ed. Adelaide, Australia: Springer Berlin Heidelberg, 2011, pp. 435–444.

[26] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, 2009, pp. 190–199.

[27] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in Proceedings of the 5th annual international conference on Systems documentation, 1986, pp. 24–26.

[28] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using Measures of Semantic Relatedness for Word Sense Disambiguation," in Computational Linguistics and Intelligent Text Processing, vol. 2588, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2003, pp. 241–257.

[29] R. Ferreira, R. D. Lins, F. Freitas, S. J. Simske, and M. Riss, "A New Sentence Similarity Assessment Measure Based on a Three-layer Sentence Representation," in Proceedings of the 2014 ACM Symposium on Document Engineering, 2014, pp. 25–34.

[30] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in Innovations in Machine Learning, Springer, 2006, pp. 137–186.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[32] C. Fellbaum, WordNet: An electronic lexical database. Dordrecht: Springer Netherlands, 1998.

[33] W. N. Francis and H. Kucera, "Brown corpus manual," Lett. to Ed., vol. 5, no. 2, p. 7, 1979.

[34] I. Atoum and C. H. Bong, "Joint Distance and Information Content Word Similarity Measure," in Soft Computing Applications and Intelligent Systems SE - 22, vol. 378, S. Noah, A. Abdullah, H. Arshad, A. Abu Bakar, Z. Othman, S. Sahran, N. Omar, and Z. Othman, Eds. Kuala Lumpur: Springer Berlin Heidelberg, 2013, pp. 257–267.

[35] L. C. Wee and S. Hassan, "Exploiting Wikipedia for Directional Inferential Text Similarity," in Fifth International Conference on Information Technology: New Generations, 2008, pp. 686–691.

[36] A. Islam, E. Milios, and V. Kešelj, "Text similarity using google tri-grams," in Advances in Artificial Intelligence, vol. 7310, L. Kosseim and D. Inkpen, Eds. Springer, 2012, pp. 312–317.

[37] N. Malandrakis, E. Iosif, and A. Potamianos, "DeepPurple: Estimating Sentence Semantic Similarity Using N-gram Regression Models and Web Snippets," 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, 2012, pp. 565–570.

[38] P. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," J. Artif. Intell. Res., vol. 37, no. 1, pp. 141–188, 2010.

[39] D. Kiela and S. Clark, "A Systematic Study of Semantic Vector Space Model Parameters," in Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), 2014, vol. 353, pp. 21–30.

[40] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), 2014, vol. 12, pp. 1532–1543.

[41] A. Lopopolo and E. van Miltenburg, "Sound-based distributional models," in Proceedings of the 11th International Conference on Computational Semantics, 2015, pp. 70–75.

[42] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis.," in nternational Joint Conference on Artificial Intelligence, 2007, vol. 7, pp. 1606–1611.

[43] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese, "UMBC EBIQUITY-CORE: Semantic textual similarity systems," in

Proceedings of the Second Joint Conference on Lexical and Computational Semantics, 2013, vol. 1, pp. 44–52.

[44] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," in Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 160–167.

[45] R. Collobert and J. Weston, "Fast semantic extraction using a novel neural network architecture," in Annual meeting-association for computational linguistics, 2007, vol. 45, no. 1, p. 560.

[46] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.

[47] M. T. Pilehvar, D. Jurgens, and R. Navigli, "Align , Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity," Proc. 51st Annu. Meet. Assoc. Comput. Linguist., pp. 1341–1351, 2013.

[48] E. Blanco and D. Moldovan, "A Semantic Logic-Based Approach to Determine Textual Similarity," Audio, Speech, Lang. Process. IEEE/ACM Trans., vol. 23, no. 4, pp. 683–693, Apr. 2015.

[49] L. Li, X. Hu, B.-Y. Hu, J. Wang, and Y.-M. Zhou, "Measuring sentence similarity from different aspects," in International Conference on Machine Learning and Cybernetics, 2009, 2009, vol. 4, pp. 2244–2249.

[50] Y. Li, H. Li, Q. Cai, and D. Han, "A novel semantic similarity measure within sentences," in Proceedings of 2012 2nd International Conference on Computer Science and Network Technology, 2012, pp. 1176–1179.

[51] G. Huang and J. Sheng, "Measuring Similarity between Sentence Fragments," in 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2012, pp. 327–330.

[52] M. A. Sultan, S. Bethard, and T. Sumner, "DLS@CU: Sentence Similarity from Word Alignment," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, vol. 2012, no. SemEval, pp. 241–246.

[53] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "PPDB: The Paraphrase Database.," in In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT '13, 2013, pp. 758–764.

[54] C. Banea, S. Hassan, M. Mohler, and R. Mihalcea, "UNT: A Supervised Synergistic Approach to Semantic Text Similarity," Proc. 6th Int. Work. Semant. Eval. conjunction with 1st Jt. Conf. Lex. Comput. Semant., pp. 635–642, 2012.

[55] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," Assoc. Adv. Artif. Intell., vol. 6, pp. 775–780, 2006.

[56] S. Hassan and R. Mihalcea, "Semantic Relatedness Using Salient Semantic Analysis.," in Proceedings of the 25th AAAI Conference on Artificial Intelligence, (AAAI 2011), 2011, pp. 884–889.

[57] Y. Liu, C. Sun, L. Lin, and X. Wang, "yiGou : A Semantic Text Similarity Computing System Based on SVM," in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, no. SemEval, pp. 80–84.

[58] F. Šarić, G. Glavaš, M. Karan, J. Šnajder, and B. D. Bašić, "Takelab: Systems for Measuring Semantic Text Similarity," First Jt. Conf. Lex. Comput. Semant., pp. 441–448, 2012.

[59] M. Lintean and V. Rus, "Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics," in Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, 2012, pp. 244–249.

[60] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," Lang. Cogn. Process., vol. 6, no. 1, pp. 1–28, 1991.

[61] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," Commun. ACM, vol. 8, no. 10, pp. 627–633, Oct. 1965.

[62] L. Finkelstein, E. Gabrilovich, and Y. Matias, "Placing search in context: the concept revisited," ACM Trans. Inf. Syst., vol. 20, no. 1, pp. 116–131, Jan. 2002.

[63] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional Semantics in Technicolor," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, 2012, pp. 136–145.

[64] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation," 2014.

[65] E. Huang, R. Socher, C. Manning, and A. Ng, "Improving Word Representations via Global Context and Multiple Word Prototypes," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, 2012, pp. 873–882.

[66] I. Atoum and A. Otoom, "Mining Software Quality from Software Reviews : Research Trends and Open Issues," Int. J. Comput. Trends Technol., vol. 31, no. 2, pp. 74–83, 2016.

[67] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.," Psychol. Rev., vol. 104, no. 2, pp. 211–240, 1997.

[68] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013, pp. 1–12.

[69] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, D. Cer, and A. Gonzalez-Agirre, "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, 2012, no. 3, pp. 385–393.

[70] I. Atoum, A. Otoom, and N. Kulathuramaiyer, "A Comprehensive Comparative Study of Word and Sentence Similarity Measures," International Journal of Computer Applications, vol. 135, no. 1. Foundation of Computer Science (FCS), NY, USA, pp. 10–17, 2016.