A Survey Based On Sentence Similarity Techniques and Approaches

Vaibhav Mishra[1], Dr. Nitesh Khilwani[2], Dr. Mohammad Muazzam[3]

[1] Mewar University, Chittorgarh ,Rajasthan, India

[2] Round Glass, Noida, Uttar Pradesh, India

[3] Mewar University , Chittorgarh , Rajasthan , India

Abstract Determining the similarity among words, sentences, passages and documents is a key component in several tasks such as (IR) information retrieval , word-sense disambiguation (WSD), short answer grading, document clustering, machine translation , answer selection in QAS and text summarization. In this paper we discusses the existing works based on text similarity by dividing them into three approaches; Character-Sequence based, GoW based and Knowledge based similarities. Apart form that , samples of combination among these similarities are also presented.

Keywords: Text Mining , Text Similarity, Natural Language Processing, Question Answering System, Semantic Similarity, Character-Sequence Based Similarity, GoW-Based Similarity, Knowledge-Driven Similarity.

1. INTRODUCTION

Text similarity methods play increasingly significant role in text associated research and uses of it in tasks like text classification, information retrieval, topic detection, topic tracking, document clustering, question answering, questions generation, short answer matching, machine translation, text summarization and others. Finding resemblance between words is a vital part of text similarity which is considered as a primary stage for larger section of text like sentence, passage and document similarities. Words could be similar either lexically or semantically. Words resemble lexically if they have similar character sequence. Words considered similar semantically if they have same meaning, but spelling wise differ to each other and can be used in the same way, used in the same context and can be said one is a type of another. Lexical similarity is presented in this paper through different Character-Sequence Based algorithms, Semantic similarity is introduced through GoW Based and Knowledge Driven algorithms. Character-Sequence based measures operate on string literals and character composition. A string (sequence of characters) metric is such a metric which measures similarity or dissimilarity among two text strings to get approximate string match/unmatch through comparison. GoW (Group of Words) Based similarity is a semantic similarity method that determines the similarity among words as per the information gained from big corpora (group of documents). Knowledge Driven similarity is such a semantic similarity measure which find out the degree of similarity among words using information extracted from semantic networks. Each similarity measure is presented in upcoming sections briefly.

Paper is organized in this manner: Section 2 presents CharacterSequence based algorithms by dividing them into character-based and term-based methods. Sections 3 and 4 introduce GoW based and knowledge Driven algorithms respectively. Section 5 introduce few combinations of similarity algorithms and finally secton 6 concludes this survey.

2. CHARACTER-SEQUENCE BASED SIMILARITY

Character Sequence similarity methods work on string sequences and character arrangements. A string (sequence of characters) metric is such a metric which measures similarity or dissimilarity among two text strings to get approximate string match/unmatch through comparison. Current survey present the most prevalent string similarity methods which were applied in SimMetrics package [1]. Fig. 1 shows, 14 algorithms have been introduced briefly; Half of them are character based and rest half are term-based distance methods.

2.1 Character-Based Similarity Measures

2.1.1 Longest Common Subsequence (LCS) algorithm undertake the resemblance between two strings based on length of connected chain of characters which exist in both Character Sequences.

2.1.2 The Levenshtein distance [2] or edit distance procedure additionally utilize the distance factor to calculate the similarity between given two text sequences. In real, this distance is tallying the basic number of operation expected to change one text sequence into other text sequence[3]. The Levenshtein distance between two text sequences a, b is given by lev a,b (|a|, |b|)

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & if\ min(i,j) = 0 \\ min\begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1 \end{cases} & otherwise \end{cases}$$

Where i, j are the indexes for words a, b respectively. Insertion , deletion, substitution operations for single character can be used. Constant time is required for this operation. Levenshtein distance similarity gives best outcome for occurrence of short string but in the event of long string cost of Levenshtein distance is equal to length of string.

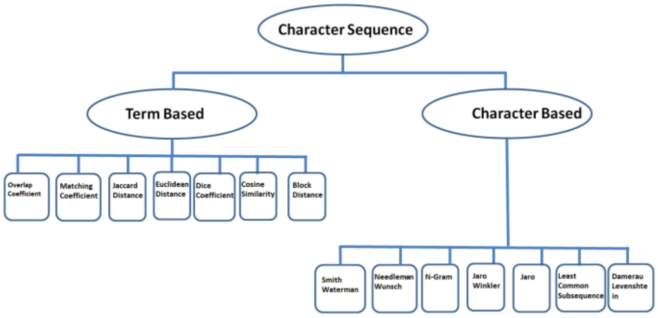

Fig 1: Character Sequence Based Similarity

2.1.3 Jaro distance is based on number and order of common characters among two strings; it takes into consideration typical spelling deviations and usually used in the area of record linkage. [4, 5].

2.1.4 Jaro–Winkler is further extension of Jaro distance; it includes a prefix scale which provide more favorable ratings to character sequence that match from beginning for a pre-defined fixed prefix length [6].

2.1.5 Needleman-Wunsch algorithm uses dynamic programming, and it is considered the first use of dynamic programming (DP) to biological sequence comparison. It applies a global alignment to find out the best arrangement over the entire two sequences. It works better on similar length sequences with a substantial degree of similarity between them [7].

2.1.6 Smith-Waterman algorithm is another example of dynamic programming (DP). It performs a local arrangement to discover the best alignment over the preserved domain of two sequences. It is considered more useful for dissimilar sequences which are supposed to comprise regions of similar sequence styles within their bigger sequence context [8].

2.1.7 N-gram technique is a sub-sequence of n characters from a given text sequence. This similarity algorithm try to match the n-grams from each character/ word in two string literals. Here distance is computed through dividing number of similar n-grams by maximum number of n-grams [9].

2.2 Term-based Similarity Measures

2.2.1 Block Distance also known as boxcar distance ,Manhattan distance, L1 distance, absolute value distance and city block distance. It find out the distance by following a grid like path where distance is travelled from one data point to the other. This distance among two items is summation of differences of their respective components [10].

2.2.2 Cosine similarity [11] is widely used approach to find the similarity between two texts. To discover the similarity between two textual contents, each part of it is represented as vector. Each word in content characterizes itself as a dimension in Euclidean space and the recurrence of each word corresponds to the dimensional value.

The Cosine similarity between two text (t1,t2)

$$SIM(t_1,t_2) = \frac{\sum_{i=1}^{n} t_{1_i} t_{2_i}}{\sqrt{\sum t_{1_i}{}^2} \times \sqrt{\sum t_{2_i}{}^2}}$$

2.2.3 Dice's coefficient is described as two times the number of common terms in the strings under comparison divided by total number of terms occurred in both of the strings [12].

2.2.4 Euclidean distance or L2 distance is calculated by calculating the squared root of squared differences sum between the two vectors elements[13].

$$\mathbf{d_{Euc}\ (\ d_1\text{-}d_2) = [(d_1\text{-}d_2).(d_1\text{-}d_2)]^{1/2}}$$

Where d1 and d2 are two vector representation of compared strings in Euclidean space.

2.2.5 Jaccard similarity Jaccard similarity identify the similarity between two usual attributes by utilizing the intersection of both then divide it by through their union [14][15]. So as per the above definitions it shows -

$$J = \frac{A_{11}}{A_{01} + A_{10} + A_{11}}$$

Where A_{11} = total number of values in binary where both vectors possess the value 1.

A01 = total number of values in binary where first vector has value 1, other has value 0.

A10 = total number of values in binary where first vector has value 0, other has value 1.

A00 = total number of values in binary where both vectors have the value 0

2.2.6 Matching Coefficient is much simpler vector based approach which counts number of similar terms, i.e. dimensions, on which both of the vectors are non zero.

2.2.7 Overlapping coefficient accept two strings a complete match if one string is a subset of another and it is similar to Dice Coefficient. The overlap coefficient also called as Szymkiewicz-Simpson coefficient is a similarity measure that is related to the Jaccard index. It deals the overlap among two sets. The measure is calculated by dividing the size of the intersection by the smaller of the size of the two sets: If set X is a subset of Y or Vice-versa, it shows that the overlap coefficient value is 1.

3. GoW BASED SIMILARITY

GoW (Group of Words) Based similarity is semantic kind of similarity measure which determines the similarity among words as per the information received from large corpora. Here GoW is a big collection of written/spoken texts which is used for linguistic research. Fig. 2 shows the GoW-Based similarity methods.

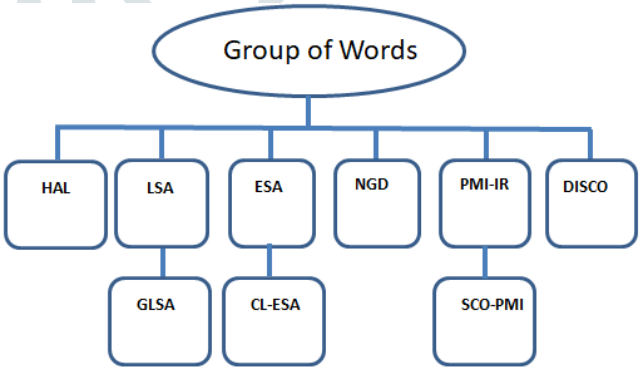

Fig 2: Group of Words Based Similarity

3.1 Hyperspace Analogue to Language (HAL) forms a semantic space [16,17] from word co-occurrences. Word-by-Word matrix is formed here with each matrix element shows the strength of association among the word denoted by row and the word denoted by column. Here the algorithm user has option to filter out small entropy columns from matrix. When the text is further analyzed, a focus word is kept at the beginning of ten word window that memorize which neighboring words should be counted as co-occurring. Matrix values are collected by weighting the inverse relationship of co-occurrence with distance from the focus word; closer neighboring words are considered to reveal focus word's semantics significantly and hence are weighted higher. HAL also keep records of word-ordering facts by handling co-occurrence differently based on if the neighboring word occurred after or before the focus word.

3.2 Latent Semantic Analysis (LSA) [18] is highly popular technique of GoW-Based similarity. LSA consider that words which are close in meaning occur in similar portions of text. Here matrix having word counts/paragraph (rows shows unique words and columns show each paragraph) is created from a big part of text. A mathematical technique called (SVD) singular value decomposition used to diminish the number of columns while keeping the similarity structure between rows. Words are finally compared by taking cosine of angle among two vectors created by any two rows.

3.3 Generalized Latent Semantic Analysis (GLSA) [19] is considered a framework aimed at computing semantically driven term and document vectors. It is an extension of LSA approach which focuses on term vectors in place of dual document-term representation. This framework requires some degree of semantic association amid terms and technique of dimensionality reduction. The GLSA approach however can combine different kind of similarity methods on space of terms having different kind of suitable method of dimensionality reduction. In last step traditional term document matrix is used which provide weights in linear combination of these term vectors.

3.4 Explicit Semantic Analysis (ESA) [20] is a method used to find out semantic relatedness among two arbitrary texts. Wikipedia-Based technique denotes terms as high-dimensional vectors; where each vector entry represents the TF-IDF weight between term and one of the Wikipedia article. The semantic relatedness between these two terms is conveyed by the cosine measure of the corresponding vectors.

3.5 The Cross-Language explicit semantic analysis (CL-ESA) [21] is ESA's multilingual generalization. CL-ESA exploits Wikipedia kind of document-aligned multilingual reference group to showcase a document as language-independent concept vector. Closeness of two documents (which are in different languages) is measured by cosine similarity among their corresponding vector representations.

3.6 Point wise Mutual Information - Information Retrieval (PMI-IR) [22] is a way of calculating the similarity among pairs of words, it applies AltaVista's Innovative Search query \ syntax to estimate probabilities. If more and more two words co-occur adjacent to each other in a web page, their PMI-IR score will also become higher.

3.7 Second-order co-occurrence point-wise mutual information (SCO-PMI) [23,24] is also semantic similarity method which uses point-wise mutual information to sort out lists of key neighbour words of two target words from a big corpus. The benefit of SOC-PMI is, it can find the similarity between those two words who don't co-occur frequently, as they co-occur with same neighbouring words.

3.8 This Normalized Google Distance (NGD) [25] is semantic similarity method resulting from number of hits given by search engine of Google for given keyword set. Keywords with similar meanings in a NLP sense tend to be "near" in Google distance units, on the other hands words with dissimilar meanings use to be farther apart.

Normalized Google Distance among two search terms x ,y is :

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log\, f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where -

- M corresponds to total number of webpages searched by Google;
- f(x) , f(y) are number of hits for terms x and y, respectively which are searched;
- f(x, y) is number of webpages on which both x and y occurred.

If two search terms x , y never co-occur together on the same webpage, but occur separately, the NGD between them is infinite. If both of the terms always co-occur together, their NGD will be zero or equivalent to the coefficient between x square and y square.

3.9 Pulling out Distributionally similar words via Co-occurrences (DISCO) [26, 27] Distributional similarity[28] among words considers that words with similar kind of meaning occur in similar kind of context. Big collections of text are statistically analyzed to find the distributional similarity. DISCO is a process that find out distributional similarity among words by applying a simple context window having size ±3 words for calculating co-occurrences.

When two words are examined for exact similarity DISCO retrieves their word vectors from the indexed data, and finds the similarity as per Lin measure [29]. If utmost distributionally similar word need to be found out; DISCO returns the 2nd order word vector for the specified word. DISCO possess two main similarity methods DISCO1 and DISCO2; DISCO1 finds the 1st order similarity among two input words depending on their collocation sets while DISCO2 finds the 2nd order similarity among two input words depending on their sets which should be of distributionally similar words.

4. KNOWLEDGE-BASED SIMILARITY

Knowledge-Based Similarity works on the concept of identifying degree of resemblance among words using information extracted form semantic [30]. WordNet [31] lexical database is most widely popular semantic network in area of estimating the Knowledge-Based similarity among words; WordNet is a big lexical database of English language. Nouns, adjectives ,verbs and adverbs are assembled into sets of cognitive synonyms called synsets. Each synset expresses a different concept. These Synsets are intertwined by means of lexical relations and conceptual-semantics.

As presented in fig. 3, Knowledge-based similarity methods can be grouped into two: methods of semantic similarity , method of semantic relatedness. Semantically similar concepts are supposed to be related through their likeness. While semantic relatedness, is a more generic idea of relatedness, not specially tied with the form or shape of the concept. In another way we can say, Semantic similarity is kind of relatedness among two words which covers a wider range of relationships among concepts which includes extra similarity relations like is-a-part-of, is-a-kind-of , is-the-opposite-of, is-a-specific-example-of [32]. There are six methods of semantic similarity; three of them are based on information content: Resnik (**res**) [33], Lin (lin) [29] and Jiang & Conrath (jcn) [34]. The other three measures are based on path length: Leacock & Chodorow (lch) [35], Wu & Palmer (wup) [36] and Path Length (path).

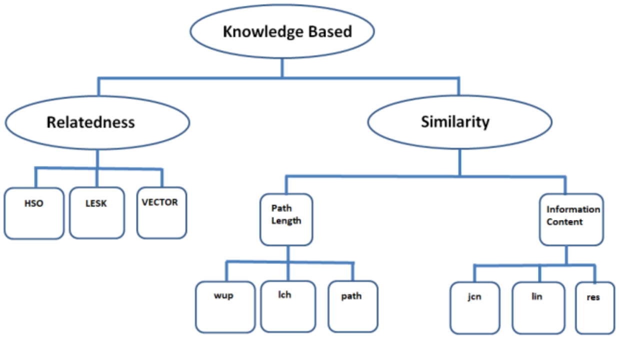

Fig 3: Knowledge Based Similarity

Related value in ***res*** method equals information content (IC) of Least Common Subsumer (which is most informative subsumer). It means value will always be either greater or equal to zero. The upper bound of value is usually quite large and differs depending on the size of corpus used to find information content values. The ***lin*** and ***jcn*** methods support the information material of Least Common Subsumer with the sum of the information material of concepts A and B themselves. The ***lin*** method scales the information material of Least Common Subsumer by using this sum, while ***jcn*** use difference of this sum in to account and the information material of Least Common Subsumer.

lch strategy restores a score which means how much comparative two word senses, depend on the shortest way that connects the senses and the greatest depth of the taxonomy wherein the senses occur**.** ***wup*** measure restores a score indicating how comparative two word senses, depend on the depth of two senses in taxonomy and their Least Common Subsumer.

path measure restores a score indicating how comparative two word senses, depend on the shortest way that associates the senses in the is-a (hypernym/hypnoym) kind of taxonomy.Besides, there are three proportions of semantic relatedness: St.Onge (hso) [37], (lesk) [38] and vector sets (vector) [39]. ***hso*** measure do work by extracting lexical chains connecting the two word senses. There are three classes of relations that are thought of: extra-solid, solid, and medium-solid. The greatest relatedness score is 16. lesk measure do work by discovering common in the glosses of the two synsets. The relatedness score is the amount of the squares of the overlapping lengths. vector measure makes a co–occurrence matrix for each given word utilized in the WordNet glosses from a designated corpus, and afterward showcase each gloss/concept through a vector which is the average of such co–occurrence vectors. In same way using average of word vectors [40] used pre-trained GloVe word vectors (from GloVe Semantic space) and their average is calculated to showcase the sentence. Cosine similarity between these vectors is used to measure similarity between these sentences.

The most prevalent packages that cover knowledge-based similarity evaluations are WordNet::Similarity1 ,Natural Language Toolkit (NLTK).

5. HYBRID-BASED SIMILARITY

Hybrid techniques utilize various similarity measures; numerous researches covered this zone. 8 semantic similarity measures were tried in [30]. 2 of these measures were GoW driven measures while other 6 were knowledge based. Initially, these 8 measures were assessed distinctly, afterward they were joined together. The best results was attained using a technique which joins a many similarity metrics into single.

A technique for estimating the semantic likeness between sentences or extremely short messages, in light of semantic and word order data[41] was introduced in [42]. To begin with, semantic likeness is calculated from a lexical KB (knowledge base) and a corpus. Second, the proposed technique thinks about the effect of word order on sentence meaningfulness. The determined word order similarity gauges the quantity of various words along with quantity of word pairs in different order.

Authors of [43] introduced a technique and called it STS (Semantic Text Similarity). This technique decides the closeness of two texts from a mix among semantic and syntactic data. They thought about 2 obligatory functions (string likeness and semantic word likeness) and a discretionary function (common-word order closeness). STS technique attained an excellent Pearson correlation coefficient for 30 sentence paired sets and beat the outcomes received in [42].

Authors of [44] introduced a methodology that consolidates GoW driven semantic similarity measure over the entire sentence alongside the knowledge based semantically likeness scores which were received for the words covered under similar syntactic roles in the two sentences. All the scores (as features) were given to AI models, like linear regression and BoW models to acquire one score which gives the level of similarity between sentences. This methodology demonstrated a huge improvement in computing the semantic likeness between sentences by the joining the knowledge based similarity measure and the GoW driven relatedness measure compare to GoW driven measure considered alone.

A Promising relationship among manual and automatic similarity outcomes were accomplished in [45] by consolidating two modules. First module computes the similarity between sentences by utilizing N-gram based similarity while the second module compute the similarity between ideas in the two sentences by utilizing concept similarity techniques and WordNet.

A framework named UKP with sensible correlation results was presented in [46], it utilized a basic log-linear regression model which is derived from training data so that it can be combine various text similarities. These measures were String relatedness , Semantic relatedness, Text extension mechanism and methods related to structure and style. The finalised UKP models comprised of a log-linear combination of around 20 features, out of 300 features developed.

Few popular datasets used for sentence similarity are like SICK (Sentences Involving Compositional Knowledge) [47] used for shared task EemEval 2014. This dataset has 10K pairs of sentences. Each pair is marked with relatedness between the sentences. This dataset is considered as standard for evaluating sentence similarity in [47]. Other datasets used are asQAsnt[49] , WikiQA[50] which are now publicly available for sentence pairing and QA domains.

6. CONCLUSION

In current survey 3 text similarity methods were talked about; Character-Sequence based , GoW based and Knowledge-based similarities. Character-Sequence based measures work on sequence of strings and character organization. 14 methods were presented; 7 of them were character based while others were term-based distance methods. GoW-Based relatedness is a semantic relatedness measure that decides the comparability between words as per information received from big corpora. 9 methods were shown; HAL,GLSA, ESA, LSA,PMI-IR, SCO-PMI, CL-ESA, DISCO and NGD. Knowledge based relatedness is one of the semantic likeness measure which is based on recognizing the degree of comparability between words utilizing information got from semantic networks. 9 methods were presented; 6 of them depended on semantic likeness - res, jcn,lin, lch, path and wup while other 3 depended on semantic relatedness - hso, vector and lesk -. Some of these methods were consolidated together in numerous researches. Lastly some useful similarity packages were referenced, for example, WordNet::Similarity , SimMetrics, and NLTK along with some popular datasets.

7. REFERENCES

[1] Chapman, S. (2006). SimMetrics : a java & c# .net library of similarity metrics, http://sourceforge.net/projects/simmetrics/.

[2] Navarro, Gonzalo, "A guided tour to approximate string matching", ACM Computing Surveys 33 (1): pp 31–88, 2001.

[3] Pradhan et. al. (2015) A Review on Text Similarity Technique used in IR and its Application, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.9

[4] Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida, Journal of the American Statistical Society, vol. 84, 406, pp 414-420.

[5] Jaro, M. A. (1995). Probabilistic linkage of large public health data file, Statistics in Medicine 14 (5-7), 491-8.

[6] Winkler W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 354–359.

[7] Needleman, B. S. & Wunsch, D. C.(1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology 48(3): 443–53.

[8] Smith, F. T. & Waterman, S. M. (1981). Identification of Common Molecular Subsequence's, Journal of Molecular Biology 147: 195–197.

[9] Alberto, B. , Paolo, R., Eneko A. & Gorka L. (2010). Plagiarism Detection across Distant Language Pairs, In Proceedings of the 23rd International Conference on Computational Linguistics, pages 37–45.

[10] Eugene F. K. (1987). Taxicab Geometry , Dover. ISBN 0-486-25202-7.

[11] Gang Qian, Shamik Sural, Yuelong Gu, Sakti Pramanik, "Similarity between euclidean and cosine angle distance for nearest neighbor queries", Proceedings of ACM Symposium on Applied Computing, 2004.

[12] Dice, L. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3).

[13] Sohangir, S., Wang, D. Improved sqrt-cosine similarity measurement. J Big Data 4, 25 (2017). https://doi.org/10.1186/s40537-017-0083-6

[14] C. Plattel, "Distributed and Incremental Clustering using Shared Nearest Neighbours," Utrecht University, 2014.

[15] Zahrotun L,(2016) Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method, Computer Engineering and Applications Vol. 5, No. 1

[16] Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. Cognitive Science Proceedings (LEA), 660-665.

[17] Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments & Computers, 28(2),203-208.

[18] Landauer, T.K. & Dumais, S.T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, 104.

[19] Matveeva, I., Levow, G., Farahat, A. & Royer, C. (2005). Generalized latent semantic analysis for term representation. In Proc. of RANLP.

[20] Gabrilovich E. & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 6–12.

[21] Martin, P., Benno, S. & Maik, A.(2008). A Wikipedia-based multilingual retrieval model. Proceedings of the 30th European Conference on IR Research (ECIR), pp. 522-530.

[22] Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning (ECML).

[23] Islam, A. and Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discov. Data 2, 2 (Jul. 2008), 1–25.

[24] Islam, A. and Inkpen, D. (2006). Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words, in Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, pp. 1033–1038.

[25] Cilibrasi, R.L. & Vitanyi, P.M.B. (2007). The Google Similarity Distance, IEEE Trans. Knowledge and Data Engineering, 19:3, 370-383.

[26] Peter, K. (2009). Experiments on the difference between semantic similarity and relatedness. In Proceedings of the 17th Nordic Conference on Computational Linguistics - NODALIDA '09, Odense, Denmark.

[27] Peter, K. (2009). Experiments on the difference between semantic similarity and relatedness. In Proceedings of the 17th Nordic Conference on Computational Linguistics - NODALIDA '09, Odense, Denmark.

[28] Ji Y, Eisenstein J. Discriminative improvements to distributional sentence similarity. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA; 2013. p. 891–96 https://www.aclweb. org/anthology/D13-1090.

[29] Lin, D. (1998b). Extracting Collocations from Text Corpora. In Workshop on Computational Terminology , Montreal, Kanada, 57–63.

[30] Mihalcea, R., Corley, C. & Strapparava, C. (2006). Corpus based and knowledge-based measures of text semantic similarity. In Proceedings of the American Association for Artificial Intelligence.(Boston, MA).

[31] Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D. & Miller, K. (1990). WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244.

[32] Patwardhan,S. , Banerjee, S. & Pedersen ,T.(2003). Using measures of semantic relatedness for word sense disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City , pp. 241–257.

[33] Resnik, R. (1995). Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada.

[34] Jiang, J. & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, Taiwan.

[35] Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet sense similarity for word sense identification. In WordNet, An Electronic Lexical Database. The MIT Press.

[36] Wu, Z.& Palmer, M. (1994). Verb semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico.

[37] Hirst, G. & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, WordNet: An electronic lexical database , pp 305–332. MIT Press.

[38] Banerjee ,S. & Pedersen, T.(2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, , Mexico City, pp 136–145.

[39] Patwardhan, V.(2003). Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, University of Minnesota, Duluth.

[40] Putra JWG, Tokunaga T. Evaluating text coherence based on semantic similarity graph. Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing; 2017. p. 76−85. https:// doi.org/10.18653/v1/W17-2410.

[41] Mamdouh Farouk. Sentence Semantic Similarity based on Word Embedding and WordNet. Proceedings of IEEE 13th International Conference on Computer Engineering and Systems; 2018. p. 33−37. https://doi.org/10.1109/ ICCES.2018.8639211.

[42] Li, Y., McLean, D., Bandar, Z., O'Shea, J., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8), 1138–1149.

[43] Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data, 2(2), 1–25.

[44] Nitish, A., Kartik, A. & Paul, B. (2012). DERI&UPM: Pushing Corpus Based Relatedness to Similarity: Shared Task System Description. First Joint Conference on Lexical and Computational Semantics (*SEM), pages 643–647, Montreal, Canada, June 7-8, 2012 Association for Computational Linguistics.

[45] Davide, B., Ronan, T., Nathalie A., & Josiane, M. (2012), IRIT: Textual Similarity Combining Conceptual Similarity with an N-Gram Comparison Method. First Joint Conference on Lexical and Computational Semantics (*SEM), pages 552–556, Montreal, Canada, June 7-8, 2012 Association for Computational Linguistics.

[46] Daniel Bar, Chris Biemann, Iryna Gurevych, and Torsten Zesch (2012), UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. First Joint Conference on Lexical and Computational Semantics (*SEM), pages 435–440, Montreal, Canada, June 7-8, 2012 Association for Computational Linguistics.

[47] Marelli M, Bentivogli L, Baroni M, Bernardi R, Menini S, Zamparelli R. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment; SemEval. 2014. https://doi.org/10.3115/v1/S14-2001. PMid: 24275290

[48] Jonas Mueller, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona; 2016. p. 2786−92. https://dl.acm.org/citation.cfm?id=3016291.

[49] Jorge Martinez-Gil, Mario Pich. Analysis of word co-occurrence in human literature for supporting semantic correspondence discovery. Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business, Graz, Austria; 2014. https://doi.org/10.1145/2637748.2638422

[50] Yang Y, Yih W, Meek C. Wikiqa: A challenge dataset for open-domain question answering. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2015. https://doi.org/10.18653/v1/D15-1237.