A Survey of Text Similarity Approaches

Wael H. Gomaa
Computer Science Department
Modern Academy for Computer Science & Management Technology
Cairo, Egypt

Aly A. Fahmy
Computer Science Department
Faculty of Computers and Information,
Cairo University
Cairo, Egypt

ABSTRACT
Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, machine translation and text summarization. This survey discusses the existing works on text similarity through partitioning them into three approaches; String-based, Corpus-based and Knowledge-based similarities. Furthermore, samples of combination between these similarities are presented.

General Terms
Text Mining, Natural Language Processing.

Keywords
Text Similarity, Semantic Similarity, String-Based Similarity, Corpus-Based Similarity, Knowledge-Based Similarity.

1. INTRODUCTION

Text similarity measures play an increasingly important role in text related research and applications in tasks such as information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization and others. Finding similarity between words is a fundamental part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities. Words can be similar in two ways lexically and semantically. Words are similar lexically if they have a similar character sequence. Words are similar semantically if they have the same thing, are opposite of each other, used in the same way, used in the same context and one is a type of another. Lexical similarity is introduced in this survey though different String-Based algorithms, Semantic similarity is introduced through Corpus-Based and Knowledge-Based algorithms. String-Based measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison. Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. Knowledge-Based similarity is a semantic similarity measure that determines the degree of similarity between words using information derived from semantic networks. The most popular for each type will be presented briefly.

This paper is organized as follows: Section two presents String-Based algorithms by partitioning them into two types character-based and term-based measures. Sections three and four introduce Corpus-Based and knowledge-Based algorithms respectively. Samples of combinations between similarity algorithms are introduced in section five and finally section six presents conclusion of the survey.

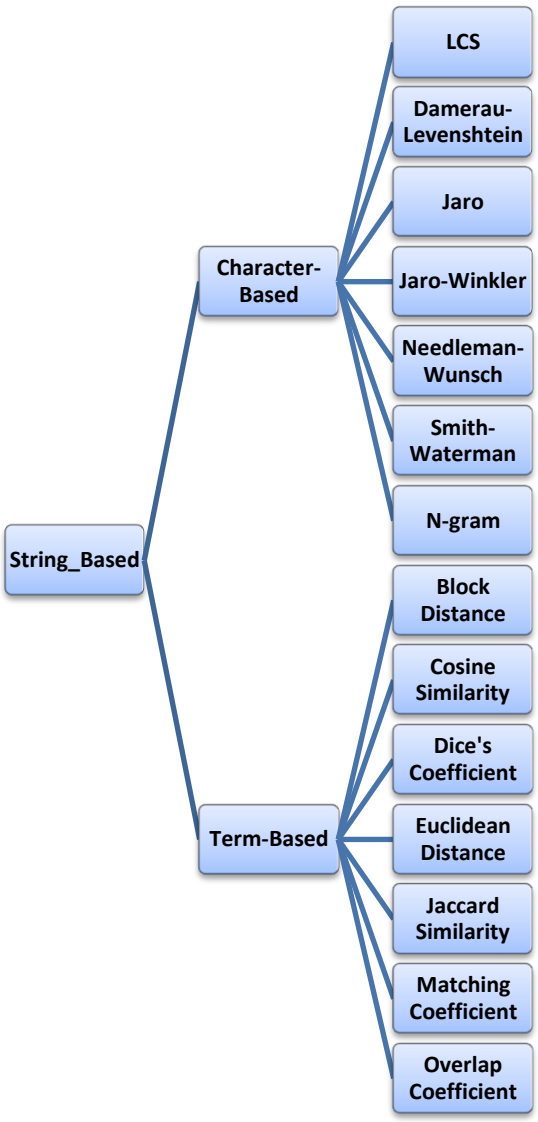

Fig 1: String-Based Similarity Measures

2. String-Based Similarity

String similarity measures operate on string sequences and character composition. A string metric is a metric that

measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison.
This survey represents the most popular string similarity measures which were implemented in SimMetrics package [1]. As shown in figure 1, fourteen algorithms will be introduced briefly; Seven of them are character based while the other are term-based distance measures.

2.1 Character-Based Similarity Measures

Longest Common SubString (LCS) algorithm considers the similarity between two strings is based on the length of contiguous chain of characters that exist in both strings.
Damerau-Levenshtein defines distance between two strings by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters [2, 3].
Jaro is based on the number and order of the common characters between two strings; it takes into account typical spelling deviations and mainly used in the area of record linkage. [4, 5].
Jaro–Winkler is an extension of Jaro distance; it uses a prefix scale which gives more favorable ratings to strings that match from the beginning for a set prefix length [6].
Needleman-Wunsch algorithm is an example of dynamic programming, and was the first application of dynamic programming to biological sequence comparison. It performs a global alignment to find the best alignment over the entire of two sequences. It is suitable when the two sequences are of similar length, with a significant degree of similarity throughout [7].
Smith-Waterman is another example of dynamic programming. It performs a local alignment to find the best alignment over the conserved domain of two sequences. It is useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context [8].
N-gram is a sub-sequence of n items from a given sequence of text. N-gram similarity algorithms compare the n-grams from each character or word in two strings. Distance is computed by dividing the number of similar n-grams by maximal number of n-grams [9].

2.2 Term-based Similarity Measures

Block Distance is also known as Manhattan distance, boxcar distance, absolute value distance, L1 distance, city block distance and Manhattan distance. It computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Block distance between two items is the sum of the differences of their corresponding components [10].
Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.
Dice's coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings [11].
Euclidean distance or L2 distance is the square root of the sum of squared differences between corresponding elements of the two vectors.
Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both strings [12].
Matching Coefficient is a very simple vector based approach which simply counts the number of similar terms, (dimensions), on which both vectors are non zero.
Overlap coefficient is similar to the Dice's coefficient, but considers two strings a full match if one is a subset of the other.

3. Corpus-Based Similarity

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. A Corpus is a large collection of written or spoken texts that is used for language research. Figure 2 shows the Corpus-Based similarity measures.
Hyperspace Analogue to Language (HAL) [13,14] creates a semantic space from word co-occurrences. A word-by-word matrix is formed with each matrix element is the strength of association between the word represented by the row and the word represented by the column. The user of the algorithm then has the option to drop out low entropy columns from the matrix. As the text is analyzed, a focus word is placed at the beginning of a ten word window that records which neighboring words are counted as co-occurring. Matrix values are accumulated by weighting the co-occurrence inversely proportional to the distance from the focus word; closer neighboring words are thought to reflect more of the focus word's semantics and so are weighted higher. HAL also records word-ordering information by treating the co-occurrence differently based on whether the neighboring word appeared before or after the focus word.
Latent Semantic Analysis (LSA) [15] is the most popular technique of Corpus-Based similarity. LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique which called singular value decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows.
Generalized Latent Semantic Analysis (GLSA) [16] is a framework for computing semantically motivated term and document vectors. It extends the LSA approach by focusing on term vectors instead of the dual document-term representation. GLSA requires a measure of semantic association between terms and a method of dimensionality reduction. The GLSA approach can combine any kind of similarity measure on the space of terms with any suitable method of dimensionality reduction. The traditional term document matrix is used in the last step to provide the weights in the linear combination of term vectors.
Explicit Semantic Analysis (ESA) [17] is a measure used to compute the semantic relatedness between two arbitrary texts. The Wikipedia-Based technique represents terms (or texts) as high- dimensional vectors; each vector entry presents the TF-IDF weight between the term and one Wikipedia article. The semantic relatedness between two terms (or texts) is expressed by the cosine measure between the corresponding vectors.

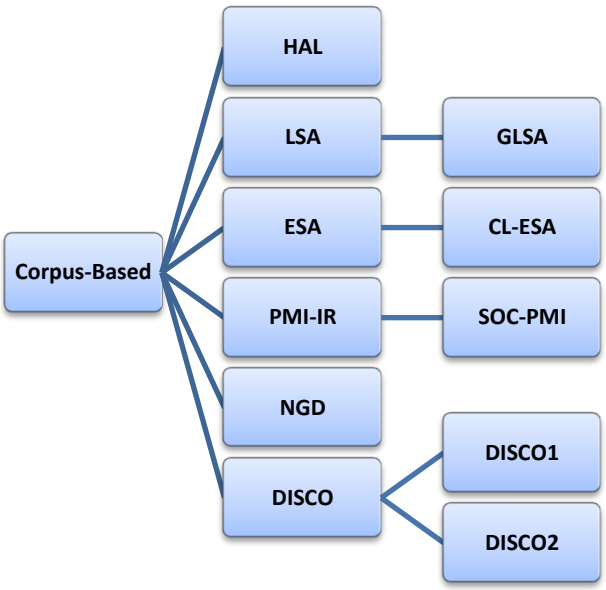

Fig 2: Corpus-Based Similarity Measures

The cross-language explicit semantic analysis (CL-ESA) [18] is a multilingual generalization of ESA. CL-ESA exploits a document-aligned multilingual reference collection such as Wikipedia to represent a document as a language-independent concept vector. The relatedness of two documents in different languages is assessed by the cosine similarity between the corresponding vector representations.

Pointwise Mutual Information - Information Retrieval (PMI-IR) [19] is a method for computing the similarity between pairs of words, it uses AltaVista's Advanced Search query \ syntax to calculate probabilities. The more often two words co-occur near each other on a web page, the higher is their PMI-IR similarity score.

Second-order co-occurrence pointwise mutual information (SCO-PMI) [20,21] is a semantic similarity measure using pointwise mutual information to sort lists of important neighbor words of the two target words from a large corpus. The advantage of using SOC-PMI is that it can calculate the similarity between two words that do not co-occur frequently, because they co-occur with the same neighboring words.

Normalized Google Distance (NGD) [22] is a semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords. Keywords with the same or similar meanings in a natural language sense tend to be "close" in units of Google distance, while words with dissimilar meanings tend to be farther apart. Specifically, the Normalized Google Distance between two search terms x and 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 is the total number of web pages searched by Google; f(x) and f(y) are the number of hits for search terms x and y, respectively; and f(x, y) is the number of web pages on which both x and y occur. If the two search terms x and y never occur together on the same web page, but do occur separately, the normalized Google distance between them is infinite. If both terms always occur together, their NGD is zero, or equivalent to the coefficient between x squared and y squared.

Extracting DIStributionally similar words using CO-occurrences (DISCO) [23, 24] Distributional similarity between words assumes that words with similar meaning occur in similar context. Large text collections are statistically analyzed to get the distributional similarity. DISCO is a method that computes distributional similarity between words by using a simple context window of size ±3 words for counting co-occurrences. When two words are subjected for exact similarity DISCO simply retrieves their word vectors from the indexed data, and computes the similarity according to Lin measure [25]. If the most distributionally similar word is required; DISCO returns the second order word vector for the given word. DISCO has two main similarity measures DISCO1 and DISCO2; DISCO1 computes the first order similarity between two input words based on their collocation sets. DISCO2 computes the second order similarity between two input words based on their sets of distributionally similar words.

4. Knowledge-Based Similarity

Knowledge-Based Similarity is one of semantic similarity measures that bases on identifying the degree of similarity between words using information derived from semantic networks [26]. WordNet [27] is the most popular semantic network in the area of measuring the Knowledge-Based similarity between words; WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

As shown in figure 3, Knowledge-based similarity measures can be divided roughly into two groups: measures of semantic similarity and measures of semantic relatedness. Semantically similar concepts are deemed to be related on the basis of their likeness. Semantic relatedness, on the other hand, is a more general notion of relatedness, not specifically tied to the shape or form of the concept. In other words, Semantic similarity is a kind of relatedness between two words, it covers a broader range of relationships between concepts that includes extra similarity relations such as is-a-kind-of, is-a-specific-example-of, is-a-part-of, is-the-opposite-of [28].

There are six measures of semantic similarity; three of them are based on information content: Resnik (***res***) [29], Lin (***lin***) [25] and Jiang & Conrath (***jcn***) [30]. The other three measures are based on path length: Leacock & Chodorow (***lch***) [31], Wu & Palmer (***wup***) [32] and Path Length (***path***).

The related value in ***res*** measure is equal to the information content (IC) of the Least Common Subsumer (most informative subsumer). This means that the value will always be greater-than or equal-to zero. The upper bound on the value is generally quite large and varies depending upon the size of the corpus used to determine information content values. The ***lin*** and ***jcn*** measures augment the information content of the Least Common Subsumer with the sum of the information content of concepts A and B themselves. The ***lin*** measure scales the information content of the Least Common Subsumer by this sum, while ***jcn*** takes the difference of this sum and the information content of the Least Common Subsumer.

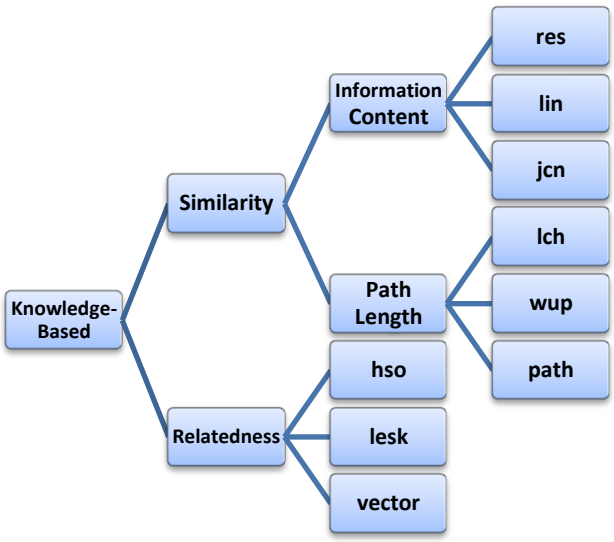

Fig 3: Knowledge-Based Similarity Measures

lch measure returns a score denoting how similar two word senses are, based on the shortest path that connects the senses and the maximum depth of the taxonomy in which the senses occur. ***wup*** measure returns a score denoting how similar two word senses are, based on the depth of the two senses in the taxonomy and that of their Least Common Subsumer.
path measure returns a score denoting how similar two word senses are, based on the shortest path that connects the senses in the is-a (hypernym/hypnoym) taxonomy.

Furthermore, there are three measures of semantic relatedness: St.Onge (***hso***) [33], Lesk (***lesk***) [34] and vector pairs (***vector***) [35]. ***hso*** measure works by finding lexical chains linking the two word senses. There are three classes of relations that are considered: extra-strong, strong, and medium-strong. The maximum relatedness score is 16. ***lesk*** measure works by finding overlaps in the glosses of the two synsets. The relatedness score is the sum of the squares of the overlap lengths. ***vector*** measure creates a co–occurrence matrix for each word used in the WordNet glosses from a given corpus, and then represents each gloss/concept with a vector that is the average of these co–occurrence vectors.
The most popular packages that cover knowledge-based similarity measures are WordNet::Similarity[1] and Natural Language Toolkit (NLTK)[2].

5. Hybrid Similarity Measures

Hybrid methods use multiple similarity measures; many researches covered this area. Eight semantic similarity measures were tested in [26]. Two of these measures were corpus-based measures and the other six were knowledge-based. Firstly, these eight algorithms were evaluated separately, then they were combined together. The best performance was achieved using a method that combines several similarity metrics into one.
A method for measuring the semantic similarity between sentences or very short texts, based on semantic and word order information was presented in [36]. First, semantic similarity is derived from a lexical knowledge base and a corpus. Second, the proposed method considers the impact of word order on sentence meaning. The derived word order similarity measures the number of different words as well as the number of word pairs in a different order.
The authors of [37] presented a method and named it Semantic Text Similarity (STS). This method determines the similarity of two texts from a combination between semantic and syntactic information. They considered two mandatory functions (string similarity and semantic word similarity) and an optional function (common-word order similarity). STS method achieved a very good Pearson correlation coefficient for 30 sentence pairs of data sets and outperformed the results obtained in [36].
The authors of [38] presented an approach that combines corpus-based semantic relatedness measure over the whole sentence along with the knowledge-based semantic similarity scores that were obtained for the words falling under the same syntactic roles in both sentences. All the scores as features were fed to machine learning models, like linear regression, and bagging models to obtain a single score giving the degree of similarity between sentences. This approach showed a significant improvement in calculating the semantic similarity between sentences by the combing the knowledge-based similarity measure and the corpus-based relatedness measure against corpus based measure taken alone.
A Promising correlation between manual and automatic similarity results were achieved in [39] by combining two modules. The first module calculates the similarity between sentences using N-gram based similarity, and the second module calculates the similarity between concepts in the two sentences using a concept similarity measure and WordNet.
A system named UKP with reasonable correlation results was introduced in [40], it used a simple log-linear regression model based on training data, to combine multiple text similarity measures. These measures were String similarity, Semantic similarity, Text expansion mechanisms and Measures related to structure and style. The UKP final models consisted of a log-linear combination of about 20 features, out of the possible 300 features implemented.

6. Conclusion

In this survey three text similarity approaches were discussed; String-based, Corpus-based and Knowledge-based similarities. String-Based measures operate on string sequences and character composition. Fourteen algorithms were introduced; Seven of them were character based while the other are term-based distance measures. Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. Nine algorithms were explained; HAL, LSA, GLSA, ESA, CL-ESA, PMI-IR, SCO-PMI, NGD and DISCO. Knowledge-Based similarity is one of semantic similarity measures that bases on identifying the degree of similarity between words using information derived from semantic networks. Nine algorithms were introduced; Six of them were based on semantic similarity -res, lin, jcn, lch, wup and path- while the other three were based on semantic relatedness -hso, lesk and vector-. Some of these algorithms were combined together in many researches. Finally useful

[1] http://wn-similarity.sourceforge.net/

[2] http://nltk.org/

similarity packages were mentioned such as SimMetrics, WordNet::Similarity and NLTK.

7. REFERENCES

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

[2] Hall , P. A. V. & Dowling, G. R. (1980) Approximate string matching, Comput. Surveys, 12:381-402.

[3] Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors, Comm. Assoc. Comput. Mach., 23:676-687.

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

[12] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547-579.

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

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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).

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

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

[26] 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).

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

AUTHORS PROFILE

Wael Hasan Gomaa, is currently working as a teacher assistant , Computer Science department , Modern Academy for Computer Science & Management Technology, Cairo, Egypt. He is a Ph.D student, Faculty of Computer and Information, Cairo University, Egypt in the field of Automatic Assessment under supervision of Prof. Aly Aly Fahmy. He received his B.Sc. and Master degrees from Faculty of Computers and Information, Helwan University, Egypt. His master thesis was entitled "Text Mining Hybrid Approach for Clustered Semantic Analysis". His research interests include Natural Language Processing, Artificial Intelligence, Data Mining and Text Mining.

Prof. Aly Aly Fahmy, is the former Dean of the Faculty of Computing and Information, Cairo University and a Professor of Artificial Intelligence and Machine Learning, in the department of Computer Science. He graduated from the Department of Computer Engineering, Technical College with honor degree. He specialized in Mathematical Logic and did his research with Dr. Hervey Gallaire, the former vice president of Xerox Global. He received a master's degree from the National School of Space and Aeronautics ENSAE, Toulouse, France, 1976 in the field of Logical Data Base systems and then obtained his PhD from the Centre for Studies and Research – CERT- DERI, 1979, Toulouse - France in the field of Artificial Intelligence.

He received practical training in the field of Operating Systems and Knowledge Based Systems in Germany and the United States of America. He participated in several national projects including the establishment of the Egyptian Universities Network (currently hosted at the Egyptian Academy of Scientific Research at the Higher Ministry of Education), building Expert Systems in the field of iron and steel industry and building Decision Support Systems for many national entities.

Prof. Fahmy's main research areas are: Data and Text Mining, Mathematical Logic, Computational Linguistics, Text Understanding and Automatic Essay Scoring and Technologies of Man- Machine Interface in Arabic. He published many refereed papers and authored the book "Decision Support Systems and Intelligent Systems" in Arabic.

He was the Director of the first Center of Excellence in Egypt in the field of Data Mining and Computer Modeling (DMCM) in the period of 2005-2010. DMCM was a virtual research center with more than 40 researchers from universities and industry. It was founded by an initiative of Dr. Tarek Kamel, Minister of Communications and Information Technology.

Prof. Aly Fahmy is currently involved in the implementation of the exploration project of Master's and Doctorate theses of Cairo University with the supervision of Prof. Dr. Hussein Khaled, Vice President of Cairo University for Postgraduate Studies and Research. The project aims to assess the practical implementation of Cairo University late strategic research plan, and to assist in the formulation of the new strategic research plan for the coming 2011 - 2015.