Short Answer Grading Using String Similarity And Corpus-Based Similarity

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*—Most automatic scoring systems use pattern based that requires a lot of hard and tedious work. These systems work in a supervised manner where predefined patterns and scoring rules are generated. This paper presents a different unsupervised approach which deals with students' answers holistically using text to text similarity. Different String-based and Corpus-based similarity measures were tested separately and then combined to achieve a maximum correlation value of 0.504. The achieved correlation is the best value achieved for unsupervised approach Bag of Words (BOW) when compared to previous work.**

Keywords-Automatic Scoring; Short Answer Grading; Semantic Similarity; String Similarity; Corpus-Based Similarity.

I. INTRODUCTION

Educational community is growing endlessly with a growing number of students, curriculums and exams. Such a growing community raised the need for scoring systems that ease the burden of scoring numerous numbers of exams and in same time guarantees the fairness of the scoring process. Automatic Scoring (AS) systems evaluate student's answer by comparing it to model answer(s). The higher correlation between student and model answers the more efficient is the scoring system. The variety in curriculums forced the AS technology to handle different kinds of students' responses, such as writing, speaking and mathematics. Writing assessment comes in two forms Automatic Essay Scoring (AES) and Short Answer Grading. Speaking assessment includes low and high entropy spoken responses while mathematical assessments include textual, numeric or graphical responses. Design and implementation of Automatic Scoring system for questions as Multiple Choice, True-False, Matching and Fill in the blank is an easy task. AS systems designed for scoring essay questions is a more difficult and complicated task as student's answers require text understanding and analysis. This paper is concerned with the automatic scoring for answers for essay questions. This research presents an unsupervised approach that deals with student's answers holistically and uses text to text similarity measures [1, 2]. The proposed model calculates the automatic score by measuring the text similarity between each word in model answer to all words in the student's answer which saves the time spent by experts to generate predefined patterns and scoring rules.

Two types of text similarity measures are presented in this research, String-based similarity, and Corpus-based similarity. String-based similarity measures operate on string sequences and character composition. Corpus-based works to identify the degree of semantic similarity between words; it depends on information derived from large corpora [3].

This paper is organized as follows:

- Section II presents related work of the main automatic short answer grading systems.
- Section III introduces the two main categories of used Similarity Algorithms.
- Section IV presents the used Data Set .
- Section V describes the proposed answer grading system.
- Section VI shows experiments' results.
- Finally, section VII presents conclusion.

II. RELATED WORK

This section describes the most famous short answer grading systems implemented for English language: C-rater [4, 5, 6], Oxford-UCLES [7, 8], Automark [9], IndusMarker [10], and Text-to-Text system [1, 2].

C-rater is the system developed by ETS, and is very reputational for high scoring accuracy for short answer responses. The reason behind high accuracy is using deep natural language processing to determine the relatedness of student response to the concepts listed in the rubric for an item. The C-rater engine applies a sequence of natural language processing steps including correcting students' spelling, determining the grammatical structure of each sentence, resolving pronoun reference, and analyzing paraphrases in the student responses [5, 6]. It has been validated on responses from multiple testing programs with different content areas including science, reading comprehension and history.

Oxford-UCLES is an information extraction short-answer scoring system that was developed at Oxford University. It uses pattern matching to evaluate the student's answers where patterns are discovered by human experts. First, it applies simple IE techniques as the nearest neighbors classification [7], then the machine learning methods like decision tree learning, Bayesian learning and inductive logic programming [8] are used.

Automarkis another system that uses IE techniques to explore the meaning or concept of text. The marking process depends mainly on content analysis in addition to specific

style features. Marking goes through 5 stages, they are discovering mark scheme templates, syntactic preprocessing, sentence analysis, pattern matching, and feedback module [9].

Indus Marker is a system that works on the structure of students' answer. It simply uses question answer markup language (QAML) to represent the required answer structures. The evaluation process starts with spell checking and some basic linguistic analysis, then the system matches the student's answer text structure with the required saved structure to compute the final mark [10].

Text-to-Text system as shown from the name; this system depends mainly on text comparison between student's answer and model answer. It doesn't use any predefined concepts or scoring rules in the evaluation process. In this approach, the evaluation process doesn't pay much attention to the subject materials, the student's answer methodology, the question type, the length of the answer and all such factors. Different semantic similarity measures were compared in [1, 2], including Knowledge-based and Corpus-based algorithms. This research depends on this approach by combining several String-based and Corpus-based similarity methods.

III. TEXT SIMILARITY MEASURES

Two categories of similarity algorithms are introduced; String-based and Corpus-based similarity .This section will handle the two measures in brief.

A. 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. Applying the concept of string metric; 13 algorithms of text similarity using Sim Metrics [11] are implemented. Six of them are character-based while the other seven are term-based distance measures

1) Character-based distance measures

Damerau-Levenshtein distance is a distance between two strings, given 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 [12,13].

Jaro algorithm 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. [14, 15].

Jaro–Winkler distance 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 [16].

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 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 [17].

Smith-Waterman algorithm is an 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 [18].

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 [19].

2) Term-based distance 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 [20].

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 [21].

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 [22].

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.

B. Corpus-Based Similarity

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora.

Latent Semantic Analysis (LSA) [23] 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.

Explicit Semantic Analysis (ESA) [24] 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 presenting 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.

Pointwise Mutual Information - Information Retrieval (PMI-IR) [25] 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 wordsco-occur near each other on a web page, the higher is their PMI-IR similarity score.

Extracting DIS tributionally similar words using CO-occurrences (DISCO [1])[26, 27]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 [28]. 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.

This research, handled the corpus-based approach via DISCO using the two main similarity measures DISCO1 and DISCO2.

IV. The Data Set

Texas[2]short answer grading data set is used [2]. It consists of ten assignments between four and seven questions each and two exams with ten questions each. These assignments/exams were assigned to an introductory computer science class at the University of North Texas. The assignments were administered as part of a Data Structures course at the University of North Texas. For each assignment, the student answers were collected via an online learning environment.

The data set as a whole contains 80 questions and 2273 student answers. The answers were scored by two human judges, using marks between 0 (completely incorrect) and 5 (perfect answer). Data set creators treated the average grade of the two evaluators as the gold standard to examine the automatic scoring task.

[1]http://www.linguatools.de/disco/disco_en.html
[2]http://lit.csci.unt.edu/index.php?P=research/downloads

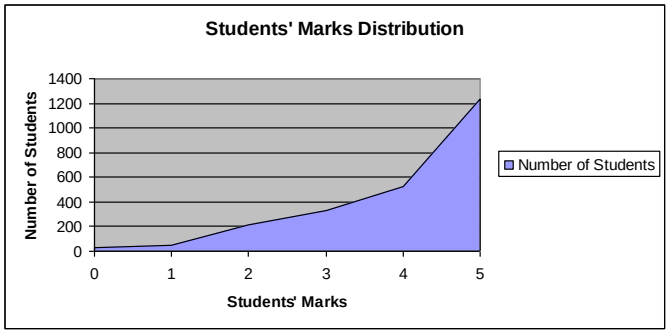

Figure 1. Students' Marks Distribuation

Table I. Sample Questions, Model Answer and Students Answers

	Question, Model Answer and Student answers	Average Grades
Question :	What is a variable?	
Model Answer :	A location in memory that can store a value.	
Student answer 1:	A variable is a location in memory where a value can be stored.	5
Student answer 2:	A named object that can hold a numerical or letter value.	4
Student answer 3:	Variable can be a integer or a string in a program	2
Question :	What is the role of a header-file?	
Model Answer :	To store a class interface, including data members and member function prototypes.	
Student answer 1:	a header file is a file used to store a list of prototype functions and data members.	5
Student answer 2:	to declare the functions being used in the classes.	3
Student answer 3:	Header files have reusable source code in a file that a programmer can use.	2.5

Figure 1 shows the students' marks distribution and table I represents a sample question, model answer, student answers and average grade.

V. Answer Grading System

Similar to all systems of automatic short answer grading, this system is based on measuring the similarity between the student's answer and the model answer to produce the final score. Then Pearson's correlation coefficient is used to specify the correlation between automatic score and average human grades.

The system goes through three stages:

The First stage is measuring the similarity between model answer and student answer using 13 String-Based algorithms previously described in section III. In this stage four methods are used to deal with strings in model and students answer; Raw, Stop, Stem, StopStem. The similarity in Raw method is computed without applying any Natural Language Processing (NLP) task. Stop Words Removing is applied in the Stop method using stop list that contains 429 words. In Stem method Porter Stemmer [29] is used to replace each non-stop with its stem without removing the stop words. Both Stop Words Removing and Stemming tasks are applied in StopStem method. Table II represents a sample of student answer using 4 methods.

Table II. Sample Student Answer with 4 Forms

Method	Student Answer
Raw	removing logical errors testing for valid data random data and actual data
Stop	removing logical errors testing valid data random data actual data
Stem	remov logic error test for valid data random data and actual data
StopStem	remov logic error test valid data random data actual data

The Second stage is measuring the similarity using DICSO1 and DISCO2 corpus-based similarity. In this stage three tasks are performed; Removing the stop words, Getting distinct words and Constructing the similarity matrix. The similarity matrix represents the similarity between each distinct word in the model answer and each distinct word in the student's answer. Each row represents one word in the model answer, and each column represents one word in the student's answer. The last two columns represent the maximum and the average similarity of each word in the model answer.

After constructing the similarity matrix, the final overall similarity is computed with two methods - Max Overall Similarity and Average Overall Similarity - by computing the average of the last two columns (Max, Average).The final overall similarity refers to the student's mark. For more clarification consider the following walkthrough example to measure the similarity between the following Model and student answer:

- Model Answer: "To store a class interface including data members and member function prototypes."
- Student Answer: "Header files have reusable source code in a file that a programmer can use."

First step is removing the stop words from the two strings ("To, a, and" in model answer and "have, in, a, that, can, use" in student answer).

Second step is getting the distinct words from the two strings; two words are considered equal if they have the same stem ("members, member" in model answer and "files, file in student answer").

Third step is constructing the similarity matrix. Table III represents the similarity matrix using DISCO2 with Wikipedia data packets.

The Third stage is combining the similarity values obtained from both string-based and corpus-based measures. Many researches adopted the idea of mixing the results from different measures to enhance the overall similarity [30, 31, 32, and 33]. The steps of the proposed combining task are illustrated in the next section.

Table III. Similarity matrix using DISCO2 Corpus-based similarity

	Header	File	reusable	Source	code	Programmer	MAX	AVG
Store	0.282	0.399	0.12	0.285	0.266	0.193	0.399	0.257
Class	0.035	0.052	0.044	0.049	0.067	0.031	0.067	0.046
Interface	0.439	0.697	0.234	0.383	0.522	0.389	0.697	0.444
Including	0.003	0.009	0.004	0.009	0.005	0.008	0.009	0.006
Data	0.468	0.61	0.163	0.017	0.45	0.253	0.61	0.326
Member	0.026	0.037	0.006	0.095	0.046	0.132	0.132	0.057
Function	0.2	0.261	0.057	0.3	0.285	0.147	0.3	0.208
prototypes	0.078	0.106	0.139	0.125	0.019	0.094	0.139	0.093
Final Overall Similarity							**0.294**	**0.179**

VI. Experiments' Results And Discussion

Pearson's correlation coefficient measure was used to specify the correlation between automatic score and average human grades.

A. Experiments Results using String-Based Similarity

As mentioned above; 13 string-based algorithms were tested with four different methods Raw, Stop, Stem and Stop Stem. Table IV represents the correlation results between model and student answer using both Character-based and Term-based measures.

In character-based distance measures; N-gram similarity got the best correlation value 0.435 applied to the raw text by mixing the results obtained from both bi-gram and tri-gram similarity measures.

Table IV. Similarity matrix using DISCO2 Corpus-based similarity

	Raw	Stop	Stem	StopStem
Character-based distance measures				
Damerau-Levenshtein	0.338	0.324	0.317	0.315
Jaro	0.144	0.229	0.146	0.205
Jaro–Winkler	0.151	0.245	0.169	0.223
Needleman-Wunsch	0.265	0.265	0.255	0.258
Smith-Waterman	0.361	0.341	0.351	0.331
N-gram (bi-gram+tri-gram)	**0.435**	0.416	0.413	0.398
Term-based distance measures				
Block Distance	0.375	**0.382**	0.34	0.291
Cosine similarity	0.376	0.377	0.344	0.308
Dice's coefficient	0.368	0.379	0.337	0.307
Euclidean distance	0.326	0.338	0.312	0.281
Jaccard similarity	0.332	0.349	0.311	0.294

Matching Coefficient	0.305	0.339	0.294	0.264
Overlap coefficient	0.374	0.368	0.336	0.286

In Term-based distance measures; Block Distance got the best correlation value 0.382 applied to the text after removing the stop words from both model and student answers.

Stop word removing task enhanced the correlation results especially in Term-based measures. Stemming process didn't enhance the results for all cases.

B. Experiments Results using Corpus-Based Similarity

Disco measures are applied by using two data packets - Wikipedia and British National Corpus (BNC); features of both are presented in table V.

Table V. Disco Data Packets

	Wikipedia	**BNC**
Packet Name	en-wikipedia-20080101	en-BNC-20080721
Packet Size	5.9 Gigabyte	1.7 Gigabyte
Number of Tokens	267 million	119 million
Number of queriable words	220,000	122,000

Table VI represents the correlation results between all the model and student answer using Disco1 and Disco2 measures. As mentioned in section V, MAX and AVG refer to Max Overall Similarity and Average Overall Similarity respectively.

Table VI. Disco Data Packets

	Wikipedia		**BNC**	
	MAX	**AVG**	**MAX**	**AVG**
Disco1	**0.465**	0.445	0.450	0.412
Disco2	0.428	0.410	0.415	0.409

In Corpus-based measures; Disco1 similarity got the best correlation value 0.475 using Wikipedia data packet and Max overall similarity method. Similarity measures using Wikipedia packet got higher correlation than BNC due to the role of the corpus size and other features shown in table V. Using Max overall similarity method clearly enhanced the correlation results in all cases in corpus-based measures.

C. Experiments Results via Combining String-Based and Corpus-Based similarity

As previously mentioned, many researches adopt the idea of mixing results from different measures to enhance the overall similarity. The proposed system combined the best algorithm for each category. The three selected measures are:

- N-gram represents character-based string similarity and is applied to raw text.
- Block Distance represents term-based string similarity and is applied to text after removing stop words.
- Disco1 using Max overall similarity which represents corpus-based similarity.

Similarity values resulting from the three measures are compared, the max and average similarity value for each student's answer are selected, and then the correlation between all students and model answers is recomputed.

The four possible combinations are represented in table VII. These cases emphasize the idea of mixing String-Based Similarity measures with the Corpus-based similarity measures to get the advantages of both. The correlation results are enhanced from the best value achieved from applying all the measures separately 0.465 to 0.504 resulting from combining N-gram and Disco1 measures.

Table VII. Correlation Results based on combining method

N-gram	**Block Distance**	**Disco1**	**MAX**	**AVG**
✗	✔	✔	0.457	0.414
✔	✗	✔	**0.504**	**0.470**
✔	✔	✗	0.411	0.394
✔	✔	✔	0.475	0.443

D. Discussion

As previously mentioned in section II; the most related research to this work was introduced in [1, 2]. The discussion here is a comparison between results from the previously related researches and results from the proposed research. Care is given to researches that deal with text as bag of words (BOW) in unsupervised way, where neither complex NLP tasks nor machine learning algorithms were applied. The dataset experimented in [1] contained 21 questions and 610 answers. LSA (BNC), LSA (Wikipedia), ESA (Wikipedia) and tf*idf were experimented. The results were 0.407, 0.428, 0.468 and 0.364 respectively. The dataset experimented in [2] was Texas dataset previously introduced in section IV .LSA, ESA and tf*idf were experimented and results were 0.328, 0.395 and 0.281 respectively.

Compared to previous results the proposed system achieved better results in most experimented cases. String-based similarity measures enhanced the correlation results if compared to simple tf*idf method. Also Disco similarity measures achieved better results than the most known Corpus-based methods LSA and ESA.

Combining String-based and Corpus-based similarity in unsupervised way raised the correlation results to 0.504. This value is the best correlation result achieved compared to other previous work and is very promising as these measures don't need any complex supervised learning task or NLP tasks such as Part of Speech and Syntax parsing. Also this value is very near to correlation values obtained from learning using different supervised machine learning algorithms and graph alignment in [2].

VII. CONCLUSION

In this research, short answer grading task is handled from an unsupervised approach which is bag of words. This approach is easy to implement as it neither requires complex NLP tasks nor supervised learning algorithms. The used data set contains 81 questions and 2273 student answers.

The proposed model goes through three stages: The First stage is measuring the similarity between model answer and student answer using 13 String-Based algorithms. Six of them were Character-based and the other seven were Term-based measures. The best correlation values achieved using Character-based and term-based were 0.435 and 0.382 using N-gram and Block distance respectively. The Second stage was measuring the similarity using DICSO1 and DISCO2 Corpus-based similarity. Disco1 achieved 0.465 correlation value using the max overall similarity.

The Third stage was measuring the similarity by combing String-based and Corpus-based measures. The best correlation value 0.504 was obtained from mixing N-gram with Disco1 similarity values. Proposed model achieved great results compared to previous works. The near future work focus on applying short answer grading to other language like Arabic. A very encouraging factor is the ability of Disco package to work with nine languages. A main obstacle for this task is the unavailability of short answer grading data sets in other language than English language.

REFERENCES

[1] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading", In Proceedings of the European Association for Computational Linguistics (EACL 2009), Athens, Greece, 2009.

[2] M. Mohler, R. Bunescu & R. Mihalcea, "Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, pp. 752–762, 2011.

[3] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based approaches to text semantic similarity", In Proceedings of the American Association for Artificial Intelligence (AAAI 2006), Boston, 2006.

[4] C. Leacock and M. Chodorow, "C-rater: Automated Scoring of Short-Answer Question", Computers and the Humanities, vol. 37, no. 4, pp. 389-405, Nov. 2003.

[5] J. Sukkarieh and S. Stoyanchev, "Automating model building in C-rater", In Proceedings of the Workshop on Applied Textual Inference, pages 6169, Suntec, Singapore, August, 2009.

[6] J. Z. Sukkarieh & J. Blackmore, "c-rater: Automatic Content Scoring for Short Constructed Responses", Proceedings of the 22nd International FLAIRS Conference, Association for the Advancement of Artificial Intelligence, 2009.

[7] J.Z. Sukkarieh, S.G. Pulman, and N. Raikes, "Auto-Marking 2: An Update on the UCLES-Oxford University research into using Computational Linguistics to Score Short, Free Text Responses", International Association of Educational Assessment, Philadephia, 2004.

[8] J. Z. Sukkarieh and S. G. Pulman, "Automatic Short Answer Marking". Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics, pp. 9-16, June 2005.

[9] T. Mitchell, T. Russel, P. Broomhead and N. Aldridge, "Towards robust computerized marking of free-text responses". Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough, UK: Loughborough University, 2002.

[10] Raheel Siddiqi, Christopher J. Harrison, and Rosheena Siddiqi,"Improving Teaching and Learning through Automated Short-Answer Marking" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 3, NO. 3, JULY-SEPTEMBER, 2010.

[11] S. Chapman, "Simmetrics: a java & c# .net library of similarity metrics", http://sourceforge.net/projects/simmetrics/, 2006.

[12] P. A. V. Hall and G. R. Dowling, "Approximate string matching, Comput. Surveys", 12:381-402 ,1980.

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

[14] M. A. Jaro, "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". Journal of the American Statistical Society 84 (406): 414–20, 1989.

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

[16] W. E. Winkler, "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, 1990.

[17] Needleman, B.Saul, and Wunsch, D. Christian , "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, 1970.

[18] Smith, F. Temple, Waterman, S. Michael , "Identification of Common Molecular Subsequences". Journal of Molecular Biology 147: 195–197, 1981.

[19] Alberto Barr´on-Cede˜no, Paolo Rosso, Eneko Agirre, and Gorka Labaka, "Plagiarism Detection across Distant Language Pairs", In Proceedings of the 23rd International Conference on Computational Linguistics, pages 37–45, 2010.

[20] Eugene F. Krause , "Taxicab Geometry" , Dover. ISBN 0-486-25202-7, 1987.

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

[22] P. Jaccard, "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, 1901.

[23] 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", Psychological Review, 104, 1997.

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

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

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

[27] Peter Kolb, "DISCO: A Multilingual Database of Distributionally Similar Words", In Proceedings of KONVENS-2008, Berlin, 2008.

[28] D. Lin, "Extracting Collocations from Text Corpora. InWorkshop on Computational Terminology" ,57–63, Montreal, Kanada,1998.

[29] M. F. Porter, "An algorithm for suffix stripping", Program, 14, 130, 1980.

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

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

[32] Nitish Aggarwal, Kartik Asooja, Paul Buitelaar. "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, Association for Computational Linguistics ,June 7-8, 2012.

[33] Davide Buscaldi, Ronan Tournier, Nathalie Aussenac-Gilles and Josiane Mothe, "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, Association for Computational Linguistics, June 7-8 , 2012.

AUTHORS PROFILE

Wael Hasan Gomaa, is currently working as a teacher assistant, Computers 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.