Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Applied Sciences

ISSN 1812-5654
DOI: 10.3923/jas.2016.209.215

Research Article
A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring

R. Mezher and N. Omar

Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia

Abstract

Background: The process of automated essays assessments is a challenging task due to the need of comprehensive evaluation in order to validate the answers accurately. The challenge increases when dealing with Arabic language where, morphology, semantic and syntactic are complex. **Methodology:** There are few research efforts have been proposed for Automatic Essays Scoring (AES) in Arabic. However, such efforts have concentrated on the semantic perspective by proposing Latent Semantic Analysis (LSA). The LSA is based on word-document co-occurrence, also called a 'Bag-of-words' approach. It is therefore blind to the syntactic information. This puts limitations on LSA's ability to capture the meaning of a sentence which depends upon both syntax and semantic. Therefore, using syntactical features may improve the process of evaluation. Hence, this study proposed a hybrid method of syntactic features and LSA for automatic essay scoring. Several pre-processing tasks have been performed in order to normalize the words with an appropriate format for processing. Then, the similarity matrix of LSA will be constructed using Term Frequency-Inverse Document Frequency (TF-IDF). After that, the cosine similarity will be carried out to identify the similarity among words. **Results:** Finally, part of speech (POS) tagging is applied in order to identify the syntactic feature of words within the similarity matrix. The dataset contains 61 questions related to environmental science with 10 answers for each question in, which the total number of answers is 610. **Conclusion:** The experimental results have shown that the syntactic feature improves the accuracy of AES for Arabic language.

Key words: Automatic essay scoring, automatic essay assessment, latent semantic analysis, syntactic, POS tagging

Received: February 16, 2016 Accepted: March 10, 2016 Published: April 15, 2016

Citation: R. Mezher and N. Omar, 2016. A hybrid method of syntactic feature and latent semantic analysis for automatic arabic essay scoring. J. Applied Sci., 16: 209-215.

Corresponding Author: N. Omar, Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Assessment plays a significant role in the educational process[1]. The interest in developing educational process by exploiting the modern technologies has been dramatically grown in the last decade. Many research efforts have been proposed in terms of producing computer-based assessment systems (CbAS). Most of these efforts have been focused on objective-type questions such as Multiple Choice Question (MCQ), multiple answer and short answer[1]. However, there is another type of questions which is subjective-type questions. Such type of question is difficult to be evaluated and requires different human assessors. This type is time-consuming and requires expertise in order to provide an accurate grade. The process of automated essays assessments is a challenging task regarding to the demand of consistent evaluation to assess the answers preciously[2].

Hence, several researchers have addressed the problem of automated essay scoring or so-called automatic essay assessment by using various techniques[3]. The key characteristic behind these techniques lies on a set of manually scored essay by human in which, the essay that intended to be assessed is compared with the pre-scored essays[4]. Usually, the manually scored essays are called pre-scored essays or training essays, whereas the essay that required to be assessed by the computer is called tested essay or automated scoring essay[5]. These terms will be substitutional used in this thesis. Such comparison aims to identify a pre-scored essay that share common features with the tested one. Hence, assigning the score of this essay to the tested essay. In fact, the comparison between the manual scored and the tested essays is performed based on specific analysis. The earliest analysis was performed based on the writing style in, which the length of paragraphs, number of sentences and number of words were the core of the comparison[6]. This approach has been criticized by many researchers due to the indifference of content analysis, which may lead to some kind of cheating[1].

Therefore, researchers have become more interested in content-based approaches in, which the lexical and semantic analysis could be performed between the manual scored and the tested essays. One of these techniques is the Latent Semantic Analysis (LSA). The LSA is a useful approach that has been commonly used in the field of automated essays assessments. Latent semantic analysis is the process of identifying semantic similarity among text sets[7]. It aims to analyse a given text using synonyms and hyponym in order to conclude the meaning of such text.

In fact, several researchers have utilized latent semantic analysis in order to assess essays answers[8,9,2,10]. The LSA is a useful approach for identifying similarity among two text set[11]. This can be performed by analysing the manually scored answers that have been provided by teachers and comparing it with the automatic answers (by the system). Several issues have been arisen in the field of automated essays answers such as treating complex languages like Arabic language. The complexity of Arabic language lies on the association between its semantic and syntactic[12]. In Arabic, in order to identify the actual meaning of certain word, it is necessary to declare its syntactic such as verb, noun, adjective and others. However, LSA does not has the ability to analyse the syntactic of a given text[13]. Therefore, this study aims to address such problem in terms of automated Arabic essay scoring.

Automatic Essay Scoring (AES) is the study that has been proposed to assess the teachers by providing an automatic approach to evaluate the score of an essay[10]. In fact, there are several techniques have been used for AES where the writing style, lexical analysis, semantic analysis, syntactic analysis and probabilistic approach have been examined in terms of providing scores[14].

Kanejiya *et al.*[13] have proposed an enhanced LSA for automatic assessment of student's answers based on syntactical features. Basically, LSA is concentrated on the semantic side where the syntactic features have been denied. Hence, the performance would significantly affected if the meaning of a given sentence is associated with grammar. Therefore, the authors have proposed syntactic features with LSA including POS tagging and parser which reasonably has contributed toward improve the accuracy.

Presently, a few research efforts have been proposed in terms of assessing essay in many languages such as Loraksa and Peachavanish[10] who proposed an automatic scoring for essay in Thai language. This method has been developed based on Artificial Neural Network (ANN) and Latent Semantic Analysis (LSA). Basically, two vectors have been built in order to represent the term frequency of the essays and the corresponding human scores. These vectors have been combined with LSA in order to enrich the synonyms and hyponym. After that, these vectors will be used as a training set for ANN to classify the semantic. In the same manner, Ishioka and Kameda[8] have proposed a Japanese Essay Scoring System (JESS), which has been developed based on LSA. The proposed method has concentrated on the semantic aspect by utilizing two matrices, one for the manually scored answer and the other for the automatic answer. Apparently, the similarity will be measured between the two matrices in order to identify the score.

Eventually, Arabic essay assessment has been addressed by Reafat *et al.*[5], who have proposed a latent semantic analysis technique in order to assess free-text essay answers in Arabic. The authors have concentrated on identifying a synonym dictionary in order to measure the similarity between the manually scored and automatic answers. Then a similarity measure which is cosine similarity has been used in terms of validating the automatic answer.

Gomaa and Fahmy[15] have introduced the first benchmark of Arabic dataset for automatic scoring essays which contains 610 students answers written in Arabic language. The authors have applied several similarity measures including string-based, n-gram and corpus-based similarity measures independently and with combination. Then they have applied k-means clustering approach in order to scale the obtained similarity values.

Finally, Alghamdi *et al.*[16] have proposed a hybrid method for automatic essay scoring in Arabic language. In fact, the proposed method consists of hybrid of Latent Semantic Analysis (LSA) with specific linguistic features including stemming, number of words and number of spelling mistakes. Apparently, the hybrid is performed using semantic and the writing style where, the semantic analysis is being performed by LSA and the writing style is represented by the number of words and spelling mistakes.

MATERIALS AND METHODS

The research design of this study consists of five main phases as shown in Fig. 1. The 1st phase is the corpus collection, which is associated with the dataset that has been used for processing. Whilst, the 2nd phase is pre-processing, which is associated with the tasks of normalization, tokenization and stemming. The 3rd phase is synonym replacement, which aims to replace each word with its corresponding synonym in order to enhance the process of identifying semanticsimilarity. For this purpose, a domain-specific dictionary has been created. The 4th phase aims to carry out the Latent Semantic Analysis (LSA). In fact, this phase consists of two sub-phases: The 1st is the Latent Semantic Analysis (LSA) which aims to produce the similarity matrix between the selected answer with the model answer. Such sub-phase aims to carry out the standard LSA using Term Frequency-Inverse Document Frequency (TF-IDF). Hence, the 2nd sub-phase aims to carry out the modified LSA where TF-IDF will be transformed into TF-POS. In fact, this modification represents the contribution of this study where LSA will be modified syntactically using POS tagging. Therefore, a comparison will be established between the standard LSA and the modified LSA where the standard LSA will be considered as a baseline. However, cosine similarity, which aims to identify the lexical similarity between the attributes within the matrix and will be applied for both standard LSA and modified LSA. Finally, the 5th phase is Evaluation, which aims to evaluate the automatic scoring compared to the manual scores.

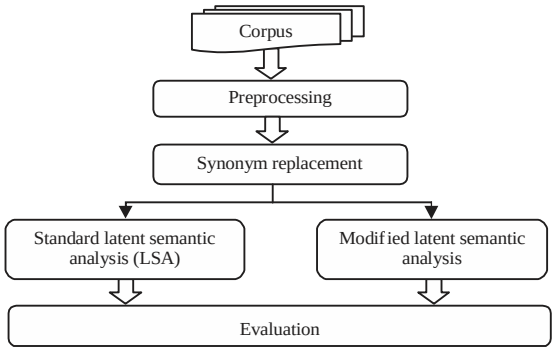

Fig. 1: Research design

Table 1: Corpus details Gomaa and Fahmy[15]

Question types	No. of questions	No. of answers
Define	18	180
Explain	6	160
What are the consequences	13	130
Why	24	240
Total	61	610

Corpus collection: The dataset used in this study is the same that introduced by Gomaa and Fahmy[15] which consists of 61 questions related to environmental science with 10 answers for each question provided by 10 students in which the total number of answers is 610. Such, questions consist of four types including "Define, explain, what are the consequences and why". Table 1 shows the details of such questions.

However, each question contains model answer that would be compared with the given answer by the students. In addition, the manual scores by the teachers have been identified in order to be compared with the automatic generated scores.

Pre-processing: This phase aims to turn the data into a suitable form by normalizing, tokenizing and stemming the data. Normalization aims to eliminate the unwanted data such as: Numbers, special characters and stopwords. According to Isa *et al.*[17], the stopwords are insignificant data that can lead to inaccurate results in terms of text classification. This is due to the false indication that could be obtained by such words

for example, the stop-word "The" could occur frequently but at the same time, such frequency does not yield valuable indication[17]. Arabic stop-words are various and can be formed with many variations. Alajmi *et al.*[18] identified several types of Arabic stop-words. Hence, this study aims to utilize the list of Arabic stop-words with its forms that have been generated by such study.

Synonym replacement: This phase aims to replace all the words with their corresponding synonyms. According to Buckeridge and Sutcliffe[7] Latent Semantic Analysis (LSA) rely mainly on the term frequency. However, there are many cases that two words with similar meaning could occur frequently but they have different lexical forms such as 'Tool' and 'Equipment'[7]. Therefore, this study aims to construct a domain-specific dictionary in order to list all the words with its potential synonyms. Hence, replacing each word with its existing synonyms in order to unify the corresponding words. Such approach has been introduced by Runeson *et al.*[19]. This can significantly provide accurate results of matching among answers where students frequently use alternative words. Note that, such domain specific dictionary has been created using Arabic WordNet (AWN)[20].

Latent semantic analysis: Latent semantic analysis is an approach that widely used in Natural Language Processing (NLP) for identifying the similarity between two groups of text[21]. It aims to analyse the relationships among two set of documents by creating a vector space for the semantic of words, terms and concepts that occurred in both documents. This can be performed by vectoring the words into two rows and columns where the words are represented in the rows and the documents represented in the columns. Then using the theory of words frequency, LSA can identify important relationship by counting the frequency of words[14]. Figure 2 shows the framework of such standard LSA.

Consider three answers and A_1, A_2 and A_3, which have sentences as shown in Table 2.

In order to represents the mentioned answers via LSA, a matrix X is being constructed in which the unique words of the three answers are represented as rows and the answers represented as columns as shown in Table 3.

Table 3 shows that the matrix will be populated based on TF-IDF in which the word frequencies will be assigned using 1 or 0 where 1 indicates the presence and 0 indicates the absence in accordance to the corresponding answer.

Considering the high dimensionality of given words, a post-processing task called Singular Value Decomposition (SVD) will be applied in order to reduce the dimensionality of the words matrix. In particular, SVD aims to decrease the number of rows without losing the similarity structure among the columns. The SVD can be calculated using the following Eq. 1:

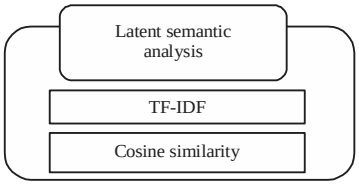

Fig. 2: Standard LSA

Table 2: Sample of answers

Contents	Meanings
A_1	Lack water affect soil erosion
A_2	Soil erosion
A_3	Overgrazing animals leads soil erosion

Table 3: LSA representation

	A_1	A_2	A_3
Abundance	1	0	0
Lack	1	0	0
Water	1	0	0
Grazing	0	0	1
Over	0	0	1
Animals	0	0	1
Affect	1	0	0
Lead	0	0	1
Erosion	0	1	1
Soil	1	1	1

$$SVD = S\Sigma U^T \quad (1)$$

where, S is the eigenvector of the product of the matrix and the transpose X^T (XX^T), U^T is the eigenvector of the product of the transpose X^T by the matrix X ($X^T X$) and Σ is the square root of eigenvalue of ($X^T X$).

After applying SVD, the cosine similarity will be computed between each pair of answers in order to identify the similarity among them. The cosine is calculated as Eq. 2:

$$\text{Cosine similarity}(A,B) = \frac{A.B}{\|A\| \times \|B\|} \quad (2)$$

where, A is the frequencies sequence of answer A_i and B is the frequencies sequence of answer A_{i+1}.

Modified LSA: The LSA uses TF-IDF in terms of representing the word's vector. Hence, LSA has a main drawback which lies in denying the syntactic aspect of the words[13]. In Arabic

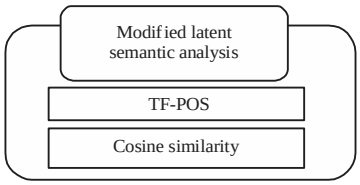

Fig. 3: Modified LSA

Table 4: TF-POS matrices

Nouns				Verbs				Adjectives			
	A_1	A_2	A_3		A_1	A_2	A_3		A_1	A_2	A_3
Water	1	0	0	Grazing	0	0	1	Lack	1	0	0
Animals	0	0	1	Affect	1	0	0	Over	0	0	1
Soil	1	1	1	Lead	0	0	1				
				Erosion	0	1	1				

language, the meaning of words is associated with the syntactic of words for instance, the word 'عصر' which yield multiple meaning including 'Afternoon', 'Era' and 'Squeeze'. Each meaning cannot be determined without using the syntactic aspect. Therefore, to identify the meaning, first it would be compulsory to identify its syntactic tag where the verb tag will refer to 'Squeeze', whereas noun tag will refer to both of 'Era' and 'Afternoon' meaning. Therefore, this study aims to propose a modified LSA using a syntactic feature. Such syntactic feature is part-of-speech (POS) tagging. The following sub-section describes it in further details.

POS tagging: Part-of-speech tagging is one of word sense disambiguation methods that aims to assign each word in certain text with a fixed set of parts of speech such as, noun, verb, adjective or adverb[22]. There are many words that have several potential tags thus, POS have been come up in order to disambiguate these words. Therefore, the main role of POS is to determine the exact tagging for each word in the corpus.

In fact, the proposed modified LSA aims to combine POS tagging with Term Frequency (TF) in order to consider the syntactic aspect in the word's vector. This has been performed by adopting TF-IDF to be TF-POS. Instead of listing all the words in the vector space, TF-POS aims to list the nouns, verbs, adjectives and others in separated vectors in respect to the answer. Figure 3 shows the framework of the proposed modified LSA.

Considering the mentioned example in LSA (three answers A_1, A_2 and A_3) which stated as follows:

A_1 = Lack of water leads to soil erosion
A_2 = Soil erosion
A_3 = Overgrazing of animals helps soil erosion

Instead of listing all the words in the vector in respect to the answers, the proposed syntactical approach TF-POS will distribute the words based on their tags. Hence, the proposed TF-POS will represent the words as in Table 4.

Table 4 shows that, the words have been divided into three vector representations based on noun, verb and adjectives. Then the Term Frequency (TF) has been identified for each word in respect to the corresponding answer. There are two main advantages of such representation. First, it can analyse the words based on the syntactical aspect whether verb, noun, adjective and others, which has a significant impact on the accuracy of automatic scoring in which each tag will be compared with its corresponding (i.e., verb with verb, noun with noun, adjective with adjective, etc.). Second, there is no need to apply the process of SVD, which requires high computational performance and complex tasks. In fact, SVD aims to reduce the dimensionality of the word's vector. Hence, the proposed TF-POS approach uses the divide and conquer strategy by utilizing multiple vectors based on the tags of words. This has the ability to reduce the dimensionality of word's vector.

Finally, cosine similarity will be applied to measure the similarity between the answers. Considering that there are multiple vectors so that, the average-cosine similarity has been used in order to identify the total average of similarity. This can be computed as Eq. 3 and 4:

$$\text{Avg} - \text{cosine}\,(A_1, A_3) = \frac{\sum \text{cosine}\,(A_1, A_3)^{\text{Noun}} + \text{cosine}\,(A_1, A_3)^{\text{Verb}} + \text{cosine}\,(A_1, A_3)^{\text{Adjective}}}{\text{No. of tags (3)}} \quad (3)$$

$$\text{Avg} - \text{cosine}\,(A_2, A_1) = \frac{\sum \text{cosine}\,(A_2, A_1)^{\text{Noun}} + \text{cosine}\,(A_2, A_1)^{\text{Verb}} + \text{cosine}\,(A_2, A_1)^{\text{Adjective}}}{\text{No. of tags (3)}} \quad (4)$$

The resulted values will be considered as the automatic score.

Evaluation: The corpus that used in this study contains the manual scores for each answer that have been performed by teachers. Therefore, the evaluation will be held by comparing the automatic scores generated by the proposed method with the manual scores. Such comparison aims to identify the similarity between the two sets in which the Root Mean Square Error (RMSE) will be computed. Due to the scores are real values, therefore euclidean distance has been used in order to calculate RMSE. In mathematics, euclidean distance is used to measure the distance between two points[23]. Therefore, the accuracy of the proposed method depends on

how the automatic and manual score is similar to each other based on euclidean distance. The accuracy can be calculated as Eq. 5:

$$RMSE = \sqrt{(S_M - S_A)^2} \quad (5)$$

where, S_M is the manual score and is the automatic score. Hence, the smaller value resulting from the previous equation, the closer distance is actually between the manual and automatic scores which leads to better accuracy of scoring.

RESULTS AND DISCUSSION

The dataset consists of answers for 10 students with a model answer in respect to 61 questions. This leads to 610 answers toward three types of questions including 'Define, explain, what are the consequences and why'. Hence, the experiments were performed based on 8 student's answers as training set (including model answer) and 3 student's answers as testing in which the evaluation is being made upon the testing set. In addition, the average values of root mean square error have been considered for all the questions. Table 5 shows the final results based on RMSE.

Table 5 shows that the results for the three students have been provided for all the answers using the standard LSA (using TF-IDF) and the modified LSA (using TF-POS). For student 1, the modified LSA has outperformed the standard LSA by achieving an average root square error of 0.2601 which is less than the average root square error for TF-IDF which is 0.2603. Although, such superiority was slightly significant however, the result of student 2 has shown a remarkable enhancement where the modified LSA has obtained an average root square error of 0.2695 compared to 0.2976 achieved by the standard LSA. As well as, the result of student 3 shown an enhancement too where the modified LSA has achieved 0.2752 of root square error compared to 0.2819 obtained by the standard LSA.

On the other hand, a comparison with a baseline should be provided in order to declare the enhancement. For this purpose, the study of Gomaa and Fahmy[15] has been addressed in which the benchmark that used in this study has been collected from such study. In fact, Gomaa and Fahmy[15] have applied multiple similarity approaches including string-based, corpus-based and knowledge-based. Since, LSA is a corpus-based similarity approach thus, the comparison will be restricted between the proposed modified LSA and the corpus-based approach used by Gomaa and Fahmy[15] which is called distributional semantic co-occurrence. Note that, the comparison is being held based on the total (i.e., average) results for all answers. Table 6 shows such results.

Table 5: Comparison among the three question types

	Student 1		Student 2		Student 3	
Question types	TF-IDF	TF-POS	TF-IDF	TF-POS	TF-IDF	TF-POS
Define	0.2273	0.2482	0.2549	0.2284	0.1833	0.2207
Explain and what are the consequences	0.2579	0.2578	0.3260	0.2859	0.2840	0.2537
Why	0.2956	0.2743	0.3120	0.2942	0.3784	0.3512
Average	0.2603	0.2601	0.2976	0.2695	0.2819	0.2752

Table 6: Comparison with baseline

Method	Root Mean Square Error (RMSE)
Gomaa and Fahmy[15] (DISCO)	0.745
Modified LSA	0.268

Table 6 shows that, there is a significant enhancement has been obtained by the proposed modified LSA compared to DISCO approach in the baseline. This can be noticed by the value of RMSE of DISCO which is 0.745 compared to the modified LSA which achieved 0.268. This out performance of LSA is due to its ability to represent the words based on frequency manner rather than the mutual information used by DISCO[24].

CONCLUSION

This study attempts to resolve the drawback of LSA which lies on the limited syntactic analysis provided by LSA. This has been performed by proposing a modified LSA for automatic essay scoring using Arabic essay answers. The modification that made for the LSA is represented by providing syntactic feature of POS tagging where the Term Frequency-Inverse Document Frequency (TF-IDF) is being transformed into TF-POS in order to analyse the syntactical class of words based on semantic analysis. Generally, this study does not address the use of machine learning. In fact, as a future direction in AEA, classifying the score would bring valuable outcomes.

ACKNOWLEDGMENT

This study is supported by the University Kebangsaan Malaysia (UKM) and funded by research grant DPP-2015-FTSM.

REFERENCES

1. Valenti, S., F. Neri and A. Cucchiarelli, 2003. An overview of current research on automated essay grading. J. Inform. Technol. Educ., 2: 319-330.
2. Landauer, T.K., 2003. Automatic essay assessment. Assess. Educ.: Principles Policy Pract., 10: 295-308.

3. Attali, Y. and J. Burstein, 2006. Automated essay scoring with e-rater® V.2. J. Technol. Learn. Assess., 4: 1-31.
4. Chung, G.K.W.K. and H.F. O'Neil Jr., 1997. Methodological approaches to online scoring of essays. CSE Technical Report No. 461, December 1997. https://www.cse.ucla.edu /products/reports/TECH461.pdf.
5. Reafat, M.M., A.A. Ewees, M.M. Eisa and A.A. Sallam, 2012. Automated assessment of students arabic free-text answers. Int. J. Cooperat. Inform. Syst., 12: 213-222.
6. Page, E.B., 2003. Project Essay Grade: PEG. In: Automated Essay Scoring: A Cross-Disciplinary Perspective, Shermis, M.D. and J. Burstein (Eds.). Lawrence Erlbaum Associates, New Jersey, pp: 43-54.
7. Buckeridge, A.M. and R.F. Sutcliffe, 2002. Disambiguating noun compounds with latent semantic indexing. Proceedings of the 2nd International Workshop on Computational Terminology, Volume 14, August 31, 2002, Taipei, Taiwan, pp: 1-7.
8. Ishioka, T. and M. Kameda, 2004. Automated Japanese essay scoring system: Jess. Proceedings of the 15th International Workshop on Database and Expert Systems Applications, August 30-September 3, 2004, Tokyo, Japan, pp: 4-8.
9. Kakkonen, T. and E. Sutinen, 2004. Automatic assessment of the content of essays based on course materials. Proceedings of the 2nd International Conference on Information Technology: Research and Education, June 28-July 1, 2004, Finland, pp: 126-130.
10. Loraksa, C. and R. Peachavanish, 2007. Automatic Thai-language essay scoring using neural network and latent semantic analysis. Proceedings of the 1st Asia International Conference on Modelling and Simulation, March 27-30, 2007, Phuket, pp: 400-402.
11. Landauer, T.K., D. Laham and P.W. Foltz, 2000. The intelligent essay assessor. IEEE Intell. Syst., 15: 27-31.
12. Ibrahim, R., Z. Eviatar and J. Aharon-Peretz, 2002. The characteristics of arabic orthography slow its processing. Neuropsychology, 16: 322-326.
13. Kanejiya, D., A. Kumar and S. Prasad, 2003. Automatic evaluation of student's answers using syntactically enhanced LSA. Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications using Natural Language Processing, Volume 2, May 31, 2003, Stroudsburg, PA., pp: 53-60.
14. Islam, M.M. and A.S.M.L. Hoque, 2010. Automated essay scoring using generalized latent semantic analysis. Proceedings of the 13th International Conference on Computer and Information Technology, December 23-25, 2010, Dhaka, pp: 358-363.
15. Gomaa, W.H. and A.A. Fahmy, 2014. Automatic scoring for answers to Arabic test questions. Comput. Speech Language, 28: 833-857.
16. Alghamdi, M., M. Alkanhal, M. Al-Badrashiny, A. Al-Qabbany, A. Areshey and A. Alharbi, 2014. A hybrid automatic scoring system for Arabic essays. J. Al Commun., 27: 103-111.
17. Isa, D., L.H. Lee, V.P. Kallimani and R. Rajkumar, 2008. Text document preprocessing with the Bayes formula for classification using the support vector machine. IEEE Trans. Knowledge Data Eng., 20: 1264-1272.
18. Alajmi, A., E.M. Saad and R.R. Darwish, 2012. Toward an ARABIC stop-words list generation. Int. J. Comput. Applic., 46: 8-13.
19. Runeson, P., M. Alexanderson and O. Nyholm, 2007. Detection of duplicate defect reports using natural language processing. Proceedings of the 29th International Conference on Software Engineering, May 20-26, 2007, Minneapolis, MN., pp: 499 -510.
20. Black, W., S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease and C. Fellbaum, 2006. Introducing the Arabic wordnet project. Proceedings of the 3rd International WordNet Conference, January 22-26, 2006, Jeju Island, South Korea, pp: 295-299.
21. Froud, H., A. Lachkar and S.A. Ouatik, 2013. Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. Int. J. Data Mining Knowledge Manage. Process, Vol. 3.
22. Navigli, R., 2009. Word sense disambiguation: A survey. ACM Comput. Surv., Vol. 41, No. 2. 10.1145/1459352. 1459355.
23. Lee, L.H., C.H. Wan, R. Rajkumar and D. Isa, 2012. An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization. Applied Intell., 37: 80-99.
24. Kolb, P., 2008. Disco: A Multilingual Database of Distributionally Similar Words. In: Textressourcen und Lexikalisches Wissen: KONVENS 2008-Erganzungsband, Storrer, A., A. Geyken, A. Siebert and K.M. Wurzner (Eds.). Berlin-Brandenburgische Akad. der Wiss., Berlin, ISBN: 9783000256110.