Journal of Computer Science 9 (10): 1369-1382, 2013
ISSN: 1549-3636

doi:10.3844/jcssp.2013.1369.1382 Published Online 9 (10) 2013 (http://www.thescipub.com/jcs.toc)

AUTOMATIC ESSAY GRADING SYSTEM FOR SHORT ANSWERS IN ENGLISH LANGUAGE

Ali Muftah Ben Omran and Mohd Juzaiddin Ab Aziz

School of Computer Science, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43600, Selangor, Malaysia

Received 2013-06-22, Revised 2013-07-25; Accepted 2013-09-02

ABSTRACT

Automatic Essay Grading (AEG) system is defined as the computer technology that evaluates and grades written prose. The short essay answer, where the essay is written in short sentences where it has two types the open ended short answer and the close ended short answer where it is our research domain based on the computer subject. The Marking of short essay answers automatically is one of the most complicated domains because it is relying heavily on the semantic similarity in meaning refers to the degree to which two sentences are similar in the meaning where both used similar words in the meaning, in this case Humans are able to easily judge if a concepts are related to each other, there for is a problem when Student use a synonym words during the answer in case they forget the target answer and they use their alternative words in the answer which will be different from the Model answer that prepared by the structure. The Standard text similarity measures perform poorly on such tasks. Short answer only provides a limited content, because the length of the text is typically short, ranging from a single word to a dozen words. This research has two propose; the first propose is Alternative Sentence Generator Method in order to generate the alternative model answer by connecting the method with the synonym dictionary. The second proposed three algorithms combined together in matching phase, Commons Words (COW), Longest Common Subsequence (LCS) and Semantic Distance (SD), these algorithms have been successfully used in many Natural Language Processing systems and have yielded efficient results. The system was manually tested on 40 questions answered by three students and evaluated by teacher in class. The proposed system has yielded %82 correlation-style with human grading, which has made the system significantly better than the other state of the art systems.

Keywords: Short Answer, COW, LCS, SD, Semantic Similarity, Synonym

1. INTRODUCTION

Automated Essay Grading (AEG) is defined as the computer technology that evaluates and scores written works (Swanson and Yamangil, 2009; Tamrakar and Dubey, 2012), AEG provides benefits to all assessment tasks' components student, evaluators and testing operation. Using AEG students can improve their writing skills by receiving a quick and useful feedback, there are two types of essays Long Essay is free text where the students are given a topic to be discussed in a long essay; it must be more than half page. Short Answers is written in short sentences or piece of text where the style is not important for marking. Short answers are typically based on the sentence length but are not required to be grammatically correct (O'Shea and Bandar, 2010). There are two types of short answer systems; the open ended system where the system able to evaluate different subjects and close ended short answer system where the system is restricted to specific subject as our proposed system. Many researchers admitted that automated grading is a highly desirable addition to the

Corresponding Author: Ali Muftah Ben Omran, School of Computer Science, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600, Selangor, Malaysia

educational tool-kit, since increased writing with feedback is known to increase the quality of student writing (Yannakoudakis *et al*., 2011). Each of those types has common features to be graded. The both types have differences in Essay Context, The Grammatical Content and the Style. There are many systems that have developed based on those features. Most of Automatic Essay Grading systems (AEG) do not require sophisticated text understanding. There are a few systems that have used to grade short essay answer (Mohler and Mihalcea, 2009). Marking short essay answers automatically is one of the most complicated domains because it is relying heavily on the semantic similarity in the meaning which is a challenging problem, since short contexts rarely share many words in common (Aziz and Ahmad, 2009). Turney and Pantel (2010) show that two words are similar to the degree that their contexts are similar; in effect showing that words that keep the same company are very similar or synonymous in meaning. From this previous work it follows that texts made up of similar words will tend to be about similar. The Standard text similarity measures perform poorly on such tasks, when Student may use a synonym words during the answer, a short answer only provides a limited context, Because the length of the text is typically short, ranging from a single word to a dozen words (Cutrone and Chang, 2011). This research is focused to build efficient automatic essay grading system for short answer in English language based on the proposed methods.

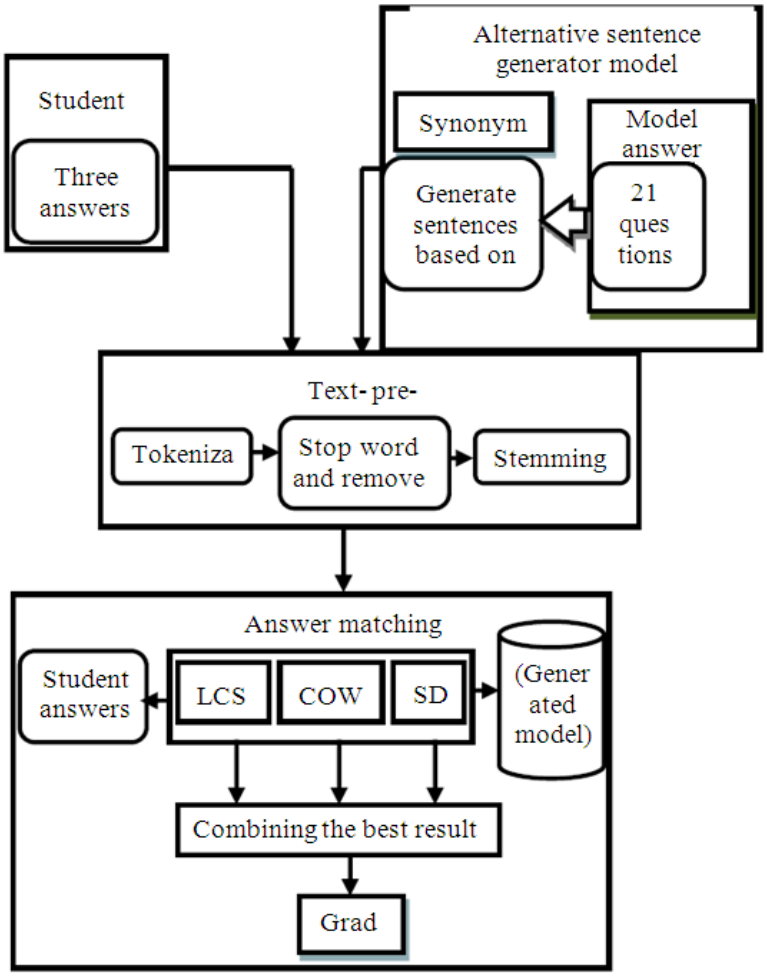

Fig. 1. Architecture of AEG system for short answer in English language

Related work: Text Similarity is a basic and important research topic in natural language processing and the similarity measure of different physical units. There are many researches that done to evaluate the student answer based on the matching between the student answer and the model answer. Research proposed by (Selvi and Bnerjee, 2010). Here they use several techniques enhanced blue method in their system, The authors claims that Keyword analysis has usually been considered a poor method, given that it is difficult to tackle problems such as synonymy or polysemy in the student answers. While on the research of Automatic Chinese Essay Scoring Using Connections between Concepts in Paragraphs proposed by Chang and Lee (2011) proposed a method which uses the similarity between the paragraphic structures in different. In research by (Shrestha, 2011) in corpus-based methods to find similarity between short text where they present a new method, based on Vector Space Model, to capture the contextual behaviour, senses and correlation, of terms and show that this method performs better than the baseline method that uses vector based cosine similarity measure. Song (2010) over his research applications of short text similarity assessment in user-interactive question answering where he has used a combined method with statistic similarity and semantic similarity. In the resent researches proposed by (Mohler and Bunescu, 2011) over Learning to grade short answer questions using semantic similarity measures and dependency graph alignments, they combine several graph alignment features with lexical semantic similarity measures using machine learning. C-rater (Sukkarieh and Blackmore, 2009) is an automated scoring engine that has been developed to score responses to content-based short answer questions. Cutrone and Chang (2011) where they evaluate student short answers based on the semantic meaning of those answers. A component-based system utilizing a Text Pre-Processing phase and a Word/Synonym Matching phase has been developed to automate the marking process. This study leverages the research conducted in recent Natural Language Processing studies to provide a fair, timely and accurate assessment of student short answers based on the semantic meaning between the model answer and the students answer.

The study is organized as follows: In section 3, we describe our proposed system Architecture in Automatic essay grading for short answers and all components associated with the system (**Fig. 1**). In section 4 we describe the evaluation of our systems and made some compression between our system and other state of the art methods and systems. Finally, in the last section, we draw some conclusions and discuss some future developments.

2. MATERIALS AND METHODS

In this research, we used general methodology to develop a Grading system of English short answer based on Alternative Sentence Generator Method and text similarity matching methods. In order to evaluate the methods for short answer grading, we have used a part the dataset proposed by (Mohler and Mihalcea, 2009). Where the total short answers in this dataset are 360 short answer (3 assignments x 40 questions/assignment x 3 student answers/question) we use part of these dataset in order to train the system and the second part to test the system. The system is containing of two main processes each process includes several techniques; the following figure shows the system.

The system is containing on several steps in order evaluate the student answer as the following.

2.1. Alternative Sentence Generator Method

In this part database of synonyms has been used to tag the synonym for each word in the model answer with their synonym, to cover all possible answers that can be used instead the original words in the model answer. The Alternative Sentence Generator Method will generate large amount of sentences for each Model Answer. Based on this operation, large amount of sentences are generated based on the number of the key words in the model answer, where the generator will take all probabilities that could be generating using the synonym. Several processes have been carried out to generate the model answer based on the synonym table. The **Fig. 2** shows the generation process.

Process separates the sentence to words (N_1, N_2, N_3,,,,,,, N_n). Taking the first word N_1 and search over the synonyms dictionary, to find their synonyms. As soon as the synonyms are found, the method will take the first synonym and replace it with the original word in the model answer and generate alternative model answer. The method will continue replacing the words with their synonym till all the probable, cases of switching are finish.

2.2. Preprocessing Phase

The Text Pre-Processing component comprises of a number of steps, which run sequentially in an effort to reduce each sentence to its canonical form.

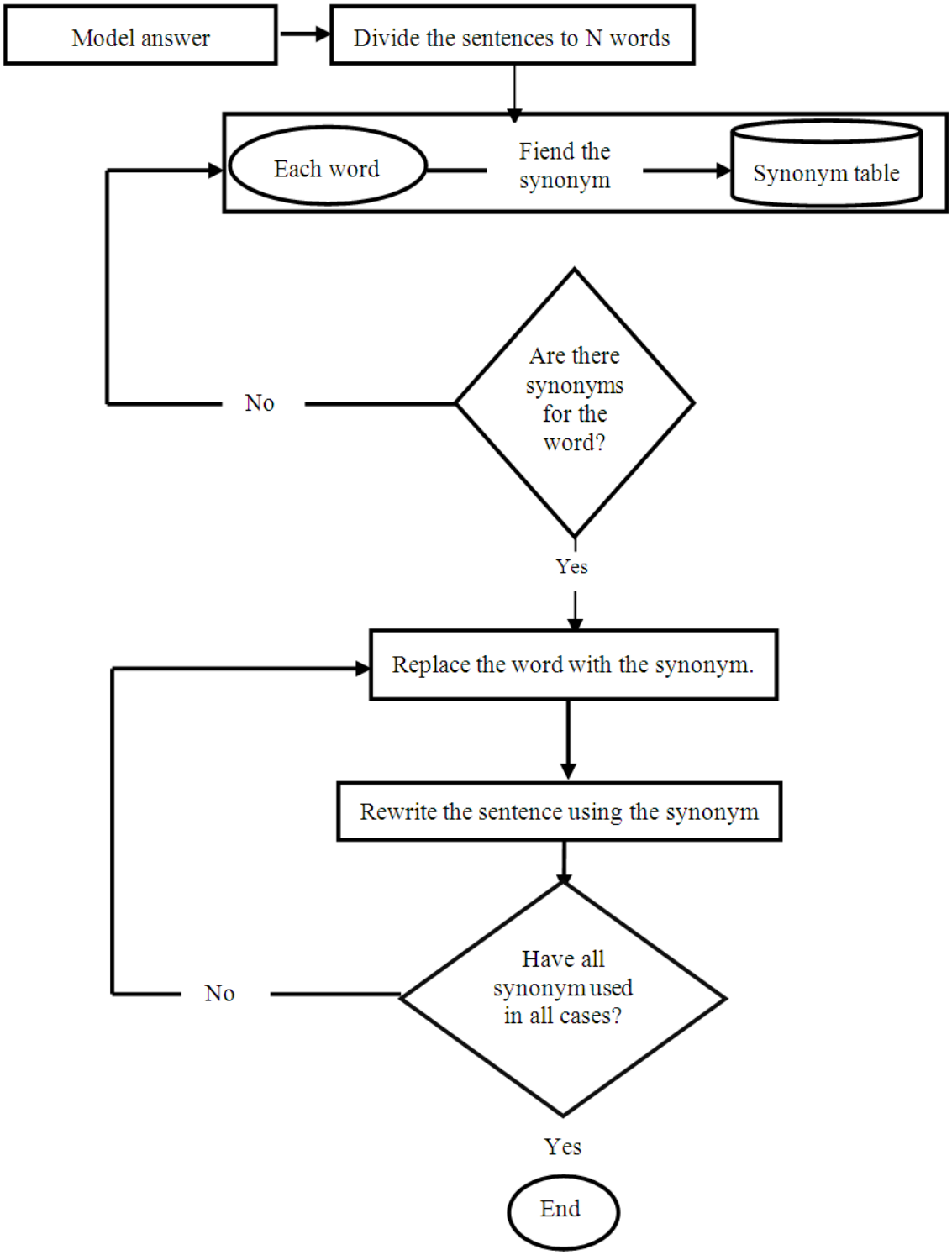

Fig. 2. Alternative sentence generator method process

These steps are applied to both, the Generated Model Answer (GMA) and the Student Answer (SA). The first step in the pre-processing is the Tokenization, where it divides the text sequence into sentences and the sentences into tokens. In alphabetic language, words are usually surrounded by whitespace. Besides the whitespace and the commas, the tokenization also removes {([\t{}():;.])} from the text and presents the words in the model answer.

In **Fig. 3** the tokenization process as it is shown in **Fig. 3** the student answer will be interned to the tokenization with 33 items includes the white spaces between the words and the words. The tekonizer will remove the white spaces between the words to 15 be ready for the next operation.

The second step in this phase is removing stop words, which include the auxiliary verbs and the preposition question words. The text is examined to determine, whether any stop words exist. The stop words, such as (if, as, the, to, at, an, a, what, where, that, on, of …..), can be found either in the generated answers or on the student answers and these stop words are removed from the text (Shrestha, 2011).

Form the example in **Fig. 4** the sentence will be cleaned for the unnecessary words as discussed earlier, this process will reduce the words from 15 to 7 important words in the answer.

After the sentence being cleaned from all stop words the next step is steaming, where the Porter Stemming algorithm, to remove all prefixes and suffixes, to get the canonical of a word. The algorithm makes a distinction between consonants and vowels in a word. Therefore, the selection of the applied rules during the stemming process is based on the sequence of consonants and vowels. The canonical form of a word is the base or lemma of that word (Turney and Pantel, 2010). For example the canonical form of the words “artist” and „artisan“ is art. In order to reduce a sentence to its canonical form, the individual words within both, the Student Answer and the Model Answer, must be examined to ensure that, they are also in their canonical form. Applying the previous results from removing the stop words to the stemming, the results will be as follows.

In **Fig. 5** the porter steamer well reduce the words in the that have gained from the previous process to their canonical form where the Porter Steamer will remove the prefixes form the words types and the parameters to their roots (type and parameter). This process is the last part in the pre-processing phase to prepare the both the student answers and the generated model answers for the matching phase.

2.3. Matching Phase

After all the possibilities of using the synonyms, in order to generate the Alternative Model Answers (AMA) and the pre-processing on those generators and the students answers are carried out. This phase investigates the use of the proposed similarity algorithms, which are Common Words (COW), Longest Common Subsequence (LCS) and Semantic Distance (SD), in the matching phase, to match the Generated Model Answer (GMA) with the student short answers. The system will run all the answers of the student on the proposed algorithms (LCS, COW and SD), where the results of all three algorithms will be combined together by giving proper weight to each algorithm based on their strengths.

2.4. Longest Common Subsequence (LCS)

The intended dataset (the Student Answer and Model Answer) will start the matching process, to find the longest subsequence common to all sequences in a set of sequences over all the student answers. It calculates the most accurate sequence by counting the letters in the sentence. The following example shows how the matching operation is done (Shan *et al*., 2009). This method works to match two of the text sequences. Using the sequence = $[y_1, y_2, y_3, ..., y_n]$ as a subsequence of another sequence X = $[x_1, x_2, x_3, ..., xm]$, if there exists a strict increasing sequence $[i_1, i_2, i_3, ..., i_n]$ of indices of such that for all j = 1, 2, 3, ...k, then $x_{ij} = y_j$. Given two sequences, X and Y, the Longest Common Subsequence (LCS) of X and Y is a common subsequence with a maximum length as in the following example.

In this example, the method calculates the longest subsequence as one whole string. In the given example the part [includename] will be the longest common subsequence in both answers. The calculation will be done using the following:

$$sim_{ICS} = \frac{2*|LCS(s_1,s_2)|}{|s_1|+|s_2|}$$

where, $|LCS(s_1, s_2)|$ is the length of the longest subsequence of s_1 and s_2.

Using the previous formula, the method calculates the longest sub sequence as one whole string. In the given example in **Fig. 6** the part [includename] will be the first part that phases the algorithm during the matching process.

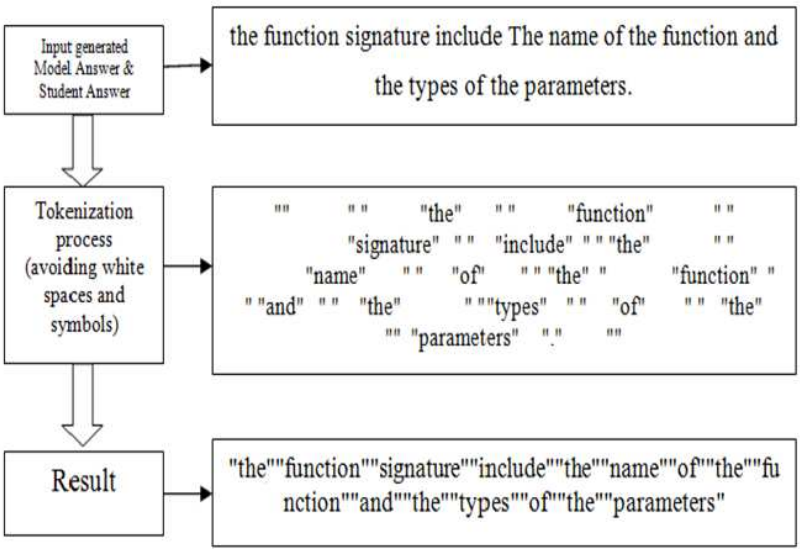

Fig. 3. Tekonization process

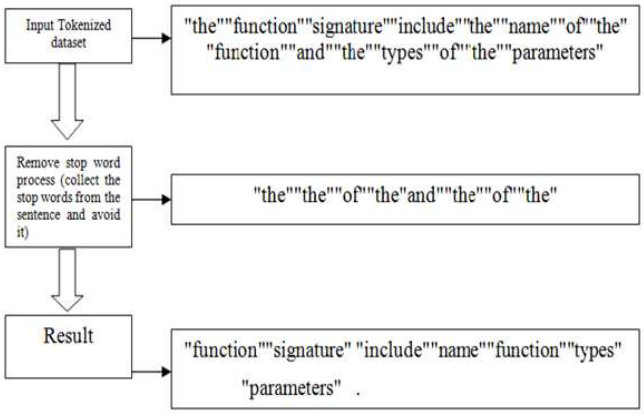

Fig. 4. Remove stop word processing

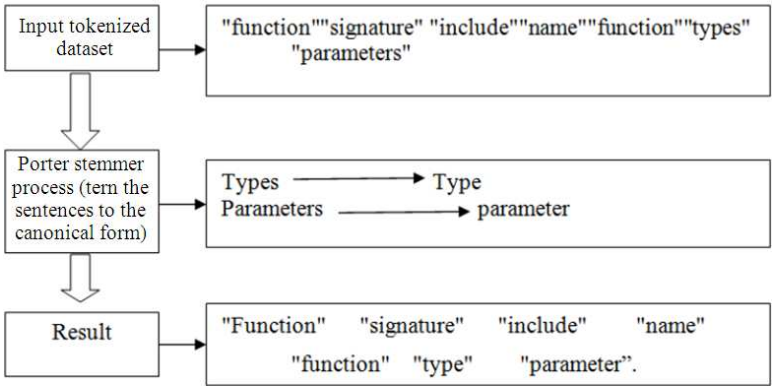

Fig. 5. Word steaming processing

Then the part [typeparameter] will be the second and the last part that can be matched between both the strings, by counting the number of common character sequence the result will be [includename]. The algorithm will calculate all possible model answer matched with the student answer and keep the best result for the best similarity to the student answer.

2.5. Common Words (COW)

It is used to match the words in both answers (the student answer and the generated model answer), where the algorithm works word by word, to determine the number of words that exist in both s_1 and s_2.

By using the following formula:

$$sim_{cow} = \frac{2*c}{|s_1| + |s_2|}$$

where, c is the number of common words between the both sentences $|s_1|$, is the total number of words in the first sentence and $|s_2|$ is the total number of words in the second sentence, the algorithm will find the best match between the generated answer and the student answer. The result will be calculated first by counting the words that are similar with c and then divide the result of 2*c on the summation of $|s_1| + |s_2|$, which are the lengths of both, the student answer and the model answer. In **Fig. 7** the algorithm will calculate all possible model answer matched with the student answer where the words (function, include, name type) the words that had matched in the given example. The algorithm will keep the best result for the best model answer that similar to the student answer.

The algorithm will calculate all possible model answer matched with the student answer and keep the best result for the best model answer that similar to the student answer.

2.6. Semantic Distance (SD)

Semantic Distance works as word by sequence, where it selects the first word from the first sentence and matches the character of this word with all the character sequences of other sentences. Here the matching process will continue for all the words in both sentences, then any two sentences, s_1, s_2, where s_1 contains the words represented as, W_{11}, $W_{12,}$…W_{1m} and s_2 contains the words represented as, W_{21}, $W_{22,}$… W_{2n}. If the word similarity occurs between W_{1i} and W_{2j}, as in the following example, the first word in the model answer will match all the character sequences for the student answer. The matching between both answers will continue with all the words as in **Fig. 8**.

For example in **Fig. 8** the word "function" will match character sequence of Model Answer in order to determine the similarity and the word "include" in **Fig. 9**. Then the matching will start from the student answer to the model answer, by matching each word in the student answer with the character sequence of the model answer.

The sentence similarity can be calculated by the following formula:

$$sim_{s\_d}(s_1, s_2) = \left(\frac{\sum_{i=1}^{m} a_i}{m} + \frac{\sum_{j=1}^{n} b_j}{n} \right) / 2$$

$$a_i = \max\{s(W_{1i}, W_{12}), s(W_{1i}, W_{22}), ..., s(W_{1i}, W_{2n})\}$$

$$b_j = \max\{s(W_{11}, W_{2j}), s(W_{12}, W_{2j}), ..., s(W_{1m}, W_{2j})\}$$

where, $\sum_{j=1}^{m} a_i$ is the summation of the number of matched characters for each word, a_i divided on the number of words in the first sentence which present the operation in **Fig. 8**. $\sum_{J=1}^{n} b_j$ is the summation of the number of matched characters for each word, $b_{j,}$ divided by the number of words in the second sentence, which present the operation in **Fig. 9**. The algorithm will calculate all possible model answer matched with the student answer and keep the best result for the best model answer that similar to the student answer.

There can be hundreds of possible answers generated using the Alternative Model Answers Generator Method, as each model has different results, during the matching. The results will be compared in order to find the generator that gains the highest result on the same model answer. The comparison between the results will be done for the three proposed methods in the matching phase, in order to find the highest mark for the student answer for each individual method. **Figure 10** shows the process of selecting best result.

Figure 10, the student selects the exam, which contains 10 questions. The first question, Q1, will be in the queue to be answered by the student. The model answer MA will be used in order to generate the Alternative Model Answers (AMA) (G1, G2, G3, G4, G5, G6…….Gn), as discussed earlier in Section (3.2.2.b). In the matching phase, the First Generation G1 will be matched with the SA over each proposed method in the matching phase and the grade is stored in S. The next generation G2 will be also matched with the SA and the result will be matched with the previous result of G1, which is stored in S. If the G2 mark > S, the result in S will be replaced with G2. As such, the student answer mark, SAM, will be the highest.

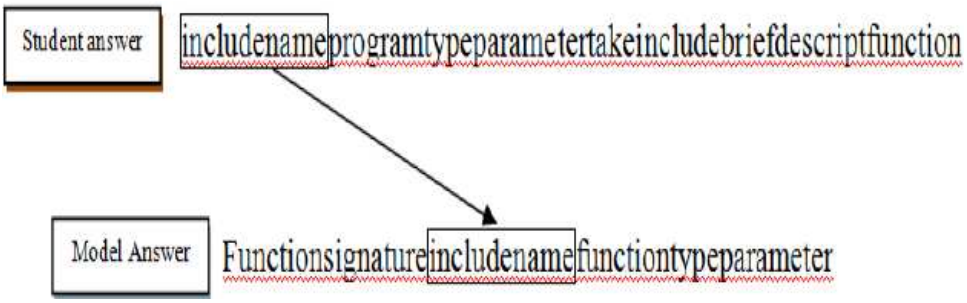

Fig. 6. Longest Common Subsequence similarity (LCS)

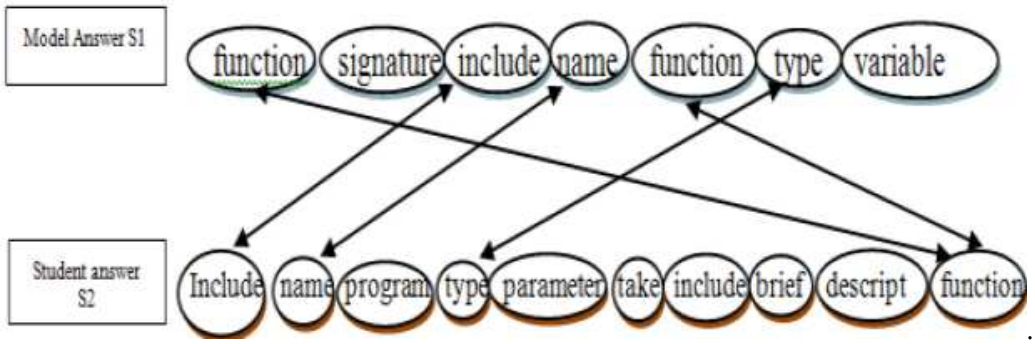

Fig. 7. Common Word Similarity (COW)

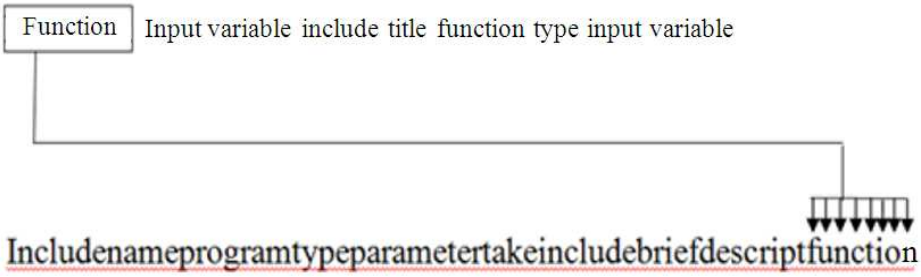

Fig. 8. SD from the SA to GMA

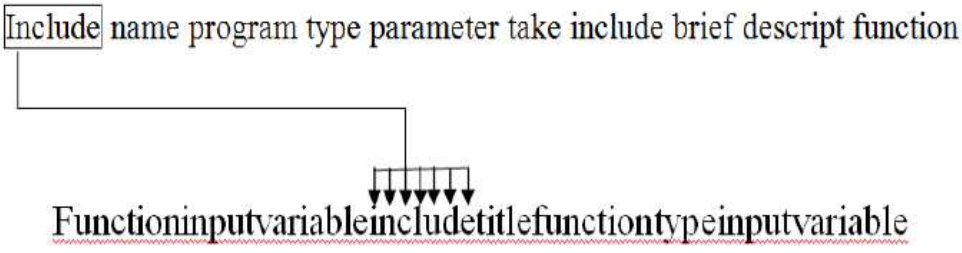

Fig. 9. SD from the GMA to SA

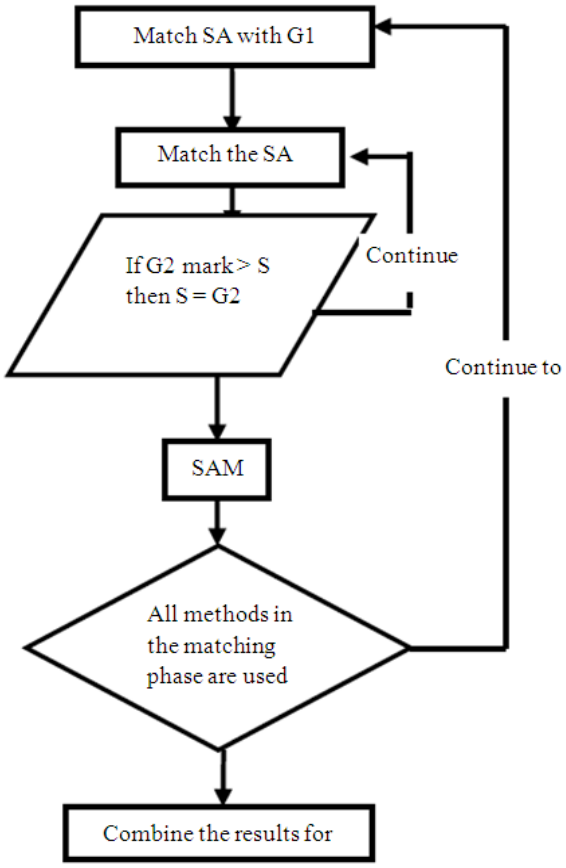

Fig. 10. Grade comparative method

2.7. The Final Result

The final result of the methods will be inserted in a combination model between the three methods, in order to get the final mark of the student's short answer, where the overall sentence similarity is calculated by weighting a smoothing factor:

$$\text{sim}(s_1, s_2) = \lambda_1 * \text{sim}_{1cs}(s_1, s_2) + \lambda_2 * \text{sim}_{cow}(s_1, s_2) + \lambda_3 * \text{sim}_{sd}(s_1, s_2)$$

where, $\text{sim}_{1cs}(s_1,s_2)$ is the result of the Longest Common Subsequence (LCS), $\text{sim}_{cow}(s_1,s_2)$ contains the result of the Common Words (COW) and $\text{sim}_{sd}(s_1,s_2)$ contains the result of the Semantic Distance (SD). λ is a weight given to each method to get the best balance in order to obtain the best result. The equation will be used to determine the best grade based on the experimental weight, λ. By giving a weight, λ, to each algorithm, the weight has been generated and tested experimentally over 100 possible attempts. The following table shows an example of how the λ is generated using the same model answer and the student answer.

From **Table 1**, it is obvious that the best generated result is G3. The combination method used the research on automated writing titled "Using latent semantic analysis to grade brief summaries: A study exploring texts at different academic levels", which was proposed by (Olmos *et al*., 2012), where the system is the effective strategy for combination. This operation of marking the questions in the exam will be continued until all the student answers for each question are matched with all the Alternative Model Answers. The total result for the entire exam will be calculated by a summation of the question results, divided by the number of questions. The weight plays important role because it makes balance between the methods to gave the best result correlated to the human, the following **Table 2** is part of the result over all the dataset that had marked by the system.

Table 1. Experimental generation for the weight

$sim(s_1,s_2)$	$\lambda_1 * sim_{lcs}(s_1,s_2)$	$\lambda_2 * sim_{cow}(s_1,s_2)$	$\lambda_3 * sim_{sd}(s_1,s_2)$	Result
G1	0.3*0.2	0.7*0.70	0.1*0.86	0.64
G2	0.5*0.2	0.2*0.70	0.3*0.86	0.49
G3	0.1*0.2	0.4*0.70	0.5*0.86	0.82
G4	0.6*0.2	0.3*0.70	0.1*0.86	0.41
G5	0.2*0.2	0.6*0.70	0.3*0.86	0.71

Table 2. Evolution results per assignment for each student

Assignment no.	Student 1	Student 3	Student 3
Assignment 1	0.987147	0.904179	0.951194
Assignment 2	0.677267	0.888022	0.803992
Assignment 3	0.783342	0.874860	0.761867

This case consist one of the most complicated cases because of the sentence length between the model answer and the student answer, the use of the method individual will gave results low according to the problem of the length form one sentence to another as LCS method results, on the other hand SD method gave a result could be higher than the human in such case. The use of the balance guaranty the result well be close to the human grade as possible.

3. RESULTS

The system produced results between 0 (not correct)-1 (full answer), Results were calculated based on comparisons in three parts; of the Human Grader and the proposed system result at the first part, in the second part between the result of the proposed system and the state of art methods with the results of our automatic grading system, to determine the level of agreement among the two assessment methods in the third part between our system result and the ASAGS system. In order to mark the student answers, three algorithms have combined together implemented, which are Common Words (COW), Longest Common Subsequence (LCS) and Semantic Distance (SD), to mark dataset, which have been graded by human grader each students" answer has graded by human grader, As the dataset has been divided into three assignments and all the assignments have been answered by three students, the evaluation in this part has been carried out, in two stages as in **Fig. 11**. The first stage has been divided into three steps. The first step has been carried over each assignment, where the assignment has been divided into three parts with each part containing seven short answers answered by one student. The second step has been executed on each assignment for all the Students and the third step has been conducted over all the assignments. The second part constitutes two parts, where the first part involves comparing the system result per assignment for all the students with the result of (Mohler and Mihalcea, 2009), where it is the same dataset that was used in order to find out the Pearson correlation in this research and the second part involves comparing with the ASAGS, which measures the correlation between the human grade and the student grade. Pearson correlation formula used, which measures the relation between them using the following equation:

$$r = \frac{\Sigma xy \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\sum y^2 - \frac{(\Sigma y)^2}{n}\right)}}$$

where in the Numerator contain the following: x is the student grade for each answer and y is the human grade for each answer, Σxy is the summation of all values for x and y multiplication in (Σx) (Σy) are the summation of x and the summation of y multiplication in each other divided N which is the number of values in both variables.

3.1. Evaluation Process

The following figure shows the evaluation process.

3.2. Correlation Measurement With Human Grade

The first stage: involves the correlation measurement for each student in each assignment where it scored (0.80-0.82) correlated to the human as in **Table 3**.

This part done for each student independently, where each student answers each assignment then the correlation between the human and the system the correlation is calculated between the human and the system marks using the pearson correlation method Discussed earlier.

The second stage: involves the measurement of the correlation style for each assignment for all the students, where each assignment has seven questions answered by three students with a total of 21 short answers, **Table 4** consist of the results in this stage.

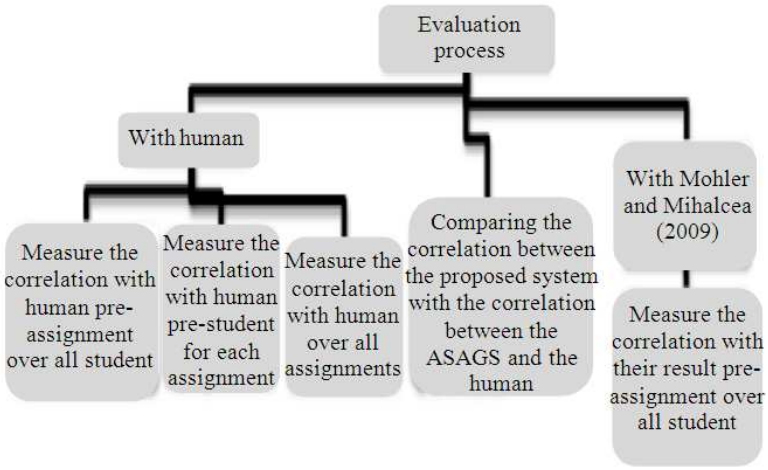

Fig. 11. Evaluation process

Table 3. The correlation between the proposed system grade and human grade

Assignment no	Parson correlation
Assignment 1	0.80
Assignment 2	0.82
Assignment 3	0.82

Table 4. The state of art method and the combined method

Techniques	Results per-assignment
Latent Semantic Analysis (LSA)	0.6465
The proposed system	
Combined method	0.82

The third stage: has been done over all students for all assignments and they got result of 82% correlated to the human.

After the Pearson correlation has calculated for each assignment, the second stage of evaluation can be done. The system evaluation can be done by comparing the proposed system evaluation form the third stage of with other systems such as the ASAGS system evaluation, where the system have use the same data set that we have used the system scored %60 correlated to the human grade (Selvi and Bnerjee, 2010). The third stage is used to compare the method with other state of the art methods which is Latent Semantic Analysis (LSA) which also used over the same dataset and scored 0.6465 correlated to the human (Mohler and Mihalcea, 2009).

3.3. Correlation Measurement with ASAGS System

The ASAGS system uses the enhanced blue method where the researchers use the same dataset in order to test the system. From the **Fig. 12** the system gave a 60% correlation to the human grader and the proposed system gave an 82% correlation to the human grader.

3.4. Correlation Measurement with Other Technique

After all the assignments have been graded, the present total for all the assignments will be calculated by a summation of the scores over the three assignments as follows:

$$\text{total} = \frac{\left(s_{1_{score}} + s_{2_{score}} + s_{3_{score}}\right)}{s_n}$$

Here $s_{1_{score}}$ is the result of the first assignment and $s_{2_{score}}$ is the result of the second assignment and $s_{3_{score}}$ is the result of the third assignment, where is the number of assignments that have been calculated, which record a 0.82 correlation with the human grade. After the total has been calculated, a comparison can be made with the work of (Mohler and Mihalcea, 2009). The research of Text-to-text Semantic Similarity for Automatic Short Answer Grading proposed by (Mohler and Mihalcea, 2009), where they have obtained the best result for LSA correlation style of about 0.6465 with the human grade per assignment for all the students, comparing with the result discussed in Chapter II, to mark the short answers in the data set that used in this research. The **Table 4** shows the techniques that have been used and their results.

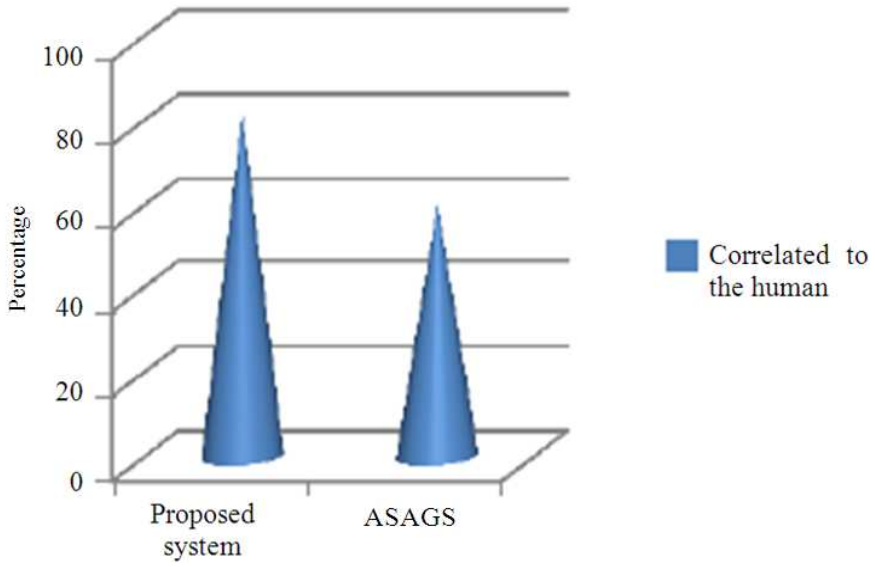

Fig. 12. Comparing the proposed system and ASAGS system with human grade

From the **Table 4**, the combination of the proposed methods they score 82% correlation with the human grade for per assignment. The method calculates the result over the answer by giving a result based on the weight for each algorithm. It is obvious the results are scalable between the algorithms from the low to height. According to all results the weight plays a major rule to produce the results in over the combination methods, where the results shows the weight worked to make balance in order to achieve the best result close to human grades.

4. DISCUSSION

The Pearson correlation style measurement method has been used in evaluation phases the following **Table 5** shows the system results.

The **Table 5** illustrates the results over all evaluations, the results shows the efficiency of the proposed system. The evaluated proposed similarity matching methods was done by comparing it with human grade, comparing it with the LSA state of the art methods that use the same dataset and comparing it with ASAGS system, where it also use the same dataset using the person correlation style measurement method. The experimental results proved that the proposed system is effective and efficient. the method such as the LSA scored lwoer result as it based on the semantic similarity (cotext similarity) deals with the senonymy and pylosmy where it is laks to the syntactic similarity mesurment (word order and the character sequence), in our approch we compined the poth aspects to overcome LSA problem wich suffers from the limitations of LSA as a bag-of-words approach.

Also our approch out perform ASAGS system based on the enhanced BLEU method where it used tommatch the student answer with limiated number of rfernces inconstract our approch provide unlimited of altrinative model answers based on the original answer according to the number of words synonymy for in the model answer.

This approch althoug is based proven to be better that the others in evalution section it still has several aspects to be improved. The grading speed may be affected by the large number of generated model answers when the algorithms of the model answers are being tested. The accuracy of the system depends on many factors such as: (i) whether there are parts that cannot be handled by the methods in the tested data, such as mathematical equations, (ii) whether the table of the synonym dictionary has been prepared well with all the words of the computer subject and (iii) whether the system has any weakness in the phrases.

Table 5. Over all evaluations

Per assignment for each student correlation with human		
Student 1	Student 2	Student 3
0.987147	0.951194	0.904179
0.677267	0.888022	0.803992
0.783342	0.87486	0.761867
Per assignment for all student correlation with human		
Assignment 1	Assignment 2	Assignment 3
0.804030601	0.823135067	0.823965202
Per assignment for all student comparing the correlation with state of the art method		
Correlation between proposed system with human	Correlations between the human and the state of the art method	
0.82	LSA 0.6465	
Overall dataset		
Correlation between Proposed system with human	Correlation between ASAGS system with human	
Correlation	%82	Correlation %60

5. CONCLUSION

This study had addressed the problem of the Essay Grading (AEG) for student short answers. The primary aims of this research was to examine the possibilities of building an Automatic Essay Grading (AEG) for short answers system based on similarity matching methods the capability of the system in successfully handling the synonym"s words of short answers matching. Much focus was given towards in two parts; firstly to the Alternative Sentence Generator Method by connecting the method with synonym dictionary in order to generate alternative model answers, secondly to the matching phases where the identified of the similarity between the student answer and the generated model answer using combination of Common Words (COW), Longest Common Subsequence (LCS) and Semantic Distance (SD) which are proposed methods over this research. The common words had been used to determine the matching between the key words in both answers. The LCS had been used to determine the common subsequence of the both answers. The SD had been used to match each word from the first answer with subsequence of the other answer and vice versa. The combination overall the methods was done by giving a proper weight for each algorithm in order to obtain the best result over the human answer. The methods were used for the matching process, which was aimed at matching the student short answers with the model answer in English language.

6. REFERENCES

Aziz, M. and F. Ahmad, 2009. Automated Marking System for Short Answer Examination (AMS-SAE). Proceedings of the IEEE Symposium on Industrial Electronics and Applications, Oct. 4-6, IEEE Xplore Press, Kuala Lumpur, pp: 47-51. DOI: 10.1109/ISIEA.2009.5356500

Chang, T. and C. Lee, 2011. Auto-Assessor: Computerized assessment system for marking student's short-answers automatically. Preceedings of the IEEE International Conference on Technology for Education, Jul. 14-16, IEEE Xplore Press, Chennai, Tamil Nadu, pp: 81-88. DOI: 10.1109/T4E.2011.21

Cutrone, L. and M. Chang, 2011. Auto-Assessor: Computerized assessment system for marking student's short-answers automatically. Proceedings of the IEEE International Conference on Technology for Education. Jul. 14-16, IEEE Xplre Press, Chennai, pp: 81-88. DOI: 10.1109/T4E.2011.21

Mohler, M. and R. Bunescu, 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (LT' 11), ACM Press, Stroudsburg, PA, USA., pp: 752-762.

Mohler, M. and R. Mihalcea, 2009. Text-to-text semantic similarity for automatic short answer grading. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, (CL' 09), ACM Press, Stroudsburg, PA, USA., pp: 567-575.

O'Shea, J. and Z. Bandar, 2010. Benchmarking short text semantic similarity. Int. J. Intell. Inform. Database Syst., 4: 103-120. DOI: 10.1504/IJIIDS.2010.032437

Olmos, R., J.A. Leon, G. Jorge-Botana and L. Escudero, 2012. Using latent semantic analysis to grade brief summaries: A study exploring texts at different academic levels. Lit Linguist Comput., 28: 388-403. DOI: 10.1093/llc/fqs065

Selvi, P. and A. Bnerjee, 2010. Automatic short-answer grading system.

Shan, J., Z. Liu and W. Zhou, 2009. Sentence similarity measure based on events and content words. Preoceedings of the Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Aug. 14-16, IEEE Xplrore Press, Tianjin, pp: 623-627. DOI: 10.1109/FSKD.2009.926

Shrestha, P., 2011. Corpus-based methods for short text similarity. Proceedings of the Montpellier, 27 Juin – 1er juillet, (JJ' 11), pp: 1-6.

Song, W., 2010. Applications of short text similarity assessment in user-interactive question answering. Thesis Ph.D., University of Hong Kong.

Sukkarieh, J.Z. and J. Blackmore, 2009. C-rater: Automatic content scoring for short constructed responses. Proceedings of the 22th International FLAIRS Conference, (FC' 09), pp: 42-44.

Swanson, B. and E. Yamangil, 2009. Correction detection and error type selection as an ESL educational aid. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (LT' 09), ACM Press, Stroudsburg, PA, USA., 357-361. Tamrakar, A. and D. Dubey, 2012. Query optimisation using natural language processing. Int. J. Tech., 3: 307-310.

Turney, P.D. and P. Pantel, 2010. From frequency to meaning: Vector space models of semantics. J. Artif. Intell. Res., 37: 141-188. DOI: 10.1613/jair.2934

Yannakoudakis, H., T. Briscoe and B. Medlock, 2011. A new dataset and method for automatically grading ESOL texts. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (LT' 11), ACM Press, Stroudsburg, PA, USA., pp: 180-189.