Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 10, No. 2, May 2019), Pages: 17-30
www.jacr.iausari.ac.ir

Presentation of an Efficient Automatic Short Answer Grading Model Based on Combination of Pseudo Relevance Feedback and Semantic Relatedness Measures

Hossein Sadr[✉1, 2], Mozhdeh Nazari Solimandarabi[2]
1) Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
2) Young Researchers and Elite Club, Rasht Branch, Islamic Azad University, Rasht, Iran
sadr@qiau.ac.ir; mozhdeh_nazary@yahoo.com

Manuscript ID: JACR-1606-1427 **Received: 2016/06/15; Accepted: 2018/02/24**

Abstract

Automatic short answer grading (ASAG) is the automated process of assessing answers based on natural language using computation methods and machine learning algorithms. Development of large-scale smart education systems on one hand and the importance of assessment as a key factor in the learning process and its confronted challenges, on the other hand, have significantly increased the need for an automated system with high flexibility for assessing exams based on texts. Generally, ASAG methods can be categorized into supervised and unsupervised approaches. Supervised approaches such as machine learning and especially deep learning methods require the manually constructed pattern. On the other hands, while in the assessment process, student's answer is compared to an ideal response and scoring is done based on their similarity, semantic relatedness and similarity measures can be considered as unsupervised approaches for this aim. Whereas unsupervised approaches do not require labeled data they are more applicable to real-world problems and are confronted with fewer limitations. Therefore, in this paper, various measures of semantic relatedness and similarity are extensively compared in the application of short answer grading. In the following, an approach is proposed for improving the performance of short answer grading systems based on semantic relatedness and similarity measures which leverages students' answers with the highest score as feedback. Empirical experiments have proved that using students' answers as feedback can considerably improve the precision of semantic relatedness and similarity measures in the automatic assessment of exams with short answers.

Keywords: *Semantic Relatedness, Short Answer Grading, Latent Semantic Analysis, Explicit Semantic Analysis, Machine Learning, Deep Learning, E-Learning System*

1. Introduction

Assessment is considered as one of the most prominent part of learning which can help us to evaluate the knowledge acquired by learners. Traditionally, assessment is done by an instructor or grader who checks students' answers and assigns them a score based on their similarity with the correct answer[1]. Using an instructor for answer grading is confronted with some limitations. One of the most crucial drawbacks refers the limited number of graders compared to the large number of learners. Furthermore, the grading process is also objective, costly and time-consuming[2]. Recent studies have

presented that there is not a high correlation between the responses assessed by different people. In fact, it can be said that the scores of students in one group with students of another group are quite different in a similar test which depends on graders' preferences and way of thinking[3,4]. According to mentioned challenges and growing use of electronic learning system, the need for an automated system for the aim of short answer grading is felt more than ever. A system which is able to evaluate answers in short time with high accuracy and eliminate the need for a grader. In such cases, employing an intelligent computer system with high speed and accuracy is essential[5, 6].

Exams can be designed in different ways such as multiple choice, true/false and fill in blanks. Whereas these kinds of questions do not require sophisticated text analysis, various intelligent and automated systems have been proposed for assessment of this kind of exams over the time[2]. Despite the flexibility and applications of these types of exams, many teachers still prefer exams based on the text. Challenge of intelligent education systems begins when short answers based on the text are used for evaluation. Considering the fact that different students have various writing styles and knowledge in answering the questions, automatic grading of these responses is considered one of the obstacles in front of intelligent education systems. As matter of fact, the aim of studies in this filed is to propose a system which is able to grade short answers based on the concept and without considering writing and spelling mistakes[7, 8].

Several studies have been concerned with automatic short answer grading in previous decades. Methods in this filed are categorized into supervised and unsupervised approaches. Although supervised approaches such as machine learning methods and recently deep learning methods have achieved considerable results in this filed, they are confronted with some limitations. In other words. Supervised approaches require labeled datasets for training and building these datasets is costly and time-consuming. Moreover, these approaches are domain specific and they may have high accuracy in training but cannot perform extremely well during prediction process. In contrast, unsupervised approaches are independent of background knowledge and therefore are more applicable to real-world problems. Since in automatic grading systems, student's answer is compared to one or several correct answers and the similarity and relatedness among them specify the score, text semantic relatedness and similarity measures can be employed as unsupervised approaches for this purpose[9, 10].

The contribution of this paper is twofold. First, various semantic relatedness and similarity measures are divided into two distinct groups of corpus-based and knowledge-based [11]and their performances are comprehensively compared in the application of short answer grading. The empirical result revealed that corpus-based measures have higher accuracy in short answer grading. Second, in order to improve the precision of semantic relatedness and similarity measure in the application of assessment, a new method is proposed which leverages students' answers with the highest score as feedback. Based on the results of experiments, using automatic feedback can significantly improve the precision of semantic related and similarity measures in the application of short answer grading.

The remainder of this paper is organized as follows: Related research in the field of automatic short answer grading are discussed in section 2. Whereas the focus of this paper is on investigating the performance of semantic relatedness and similarity measures in the application of short answer grading, the existing semantic relatedness and similarity measures are divided into two groups and comprehensively studied in

section3. The proposed method is explained in section4. Empirical experiments and corresponding results are presented in section 5. Section 6 contains the conclusion and future works.

2. Related Work

The aim of an automatic short answer grading system is to compare student answer with an ideal answer and assign a score in a specific range. Considering the fact that in large organization assessment process can be very costly and time-consuming, using an automatic method for grading and both performance increment and cost decrement is essential[12]. It is completely obvious that in a large scale education system using traditional methods for assessment is not possible. Therefore, using automatic short answer grading is considered as a necessary element (not optional) of the online training system and is considered as one of the hottest topics in the field of natural language processing, information systems and educational[13].

Short answer grading methods are generally divided into two groups. The first group contains methods which focus on correcting writing and grammatical mistakes and do not consider the concept[4]. In contrast, the second one includes methods which only focus on the concept of the answer and writing and grammatical mistakes do not have any impact on the assessment process. In this paper, the methods which only consider the concept of the answers are studied[3].

Automatic short answer grading methods which focus on the concept are categorized into two type: supervised and unsupervised approaches. The general classification of existing approaches is presented in figure 1. Supervised approaches are also divided into machine learning and manually constructed patterns. Machine learning methods employ classifiers such as Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM)[14, 15]. In recent years deep learning models have also been widely used in this filed[10]. Other of the supervised approaches also require patterns generated by an expert for comparison[16]. In other words, if the student's answer conforms to the predetermined patterns, the question is answered correctly. It must be taken into consideration that this kind of methods do not deliberate on the concept, therefore, an answer with the correct concept that do not conform to the pattern is not scored properly[17, 18]. Moreover, there are some semi-supervised methods which have higher flexibility in comparison to supervised methods. It must be noted that all of these methods require a background knowledge containing predetermined patterns and their performance has a direct dependency on them. Furthermore, constructing a rich background knowledge is really costly and time-consuming. Existing challenges and difficulties caused to propose unsupervised approaches which do not depend on the direct involvement of human resources and they have higher applicability in real-world issues[3, 19, 20].

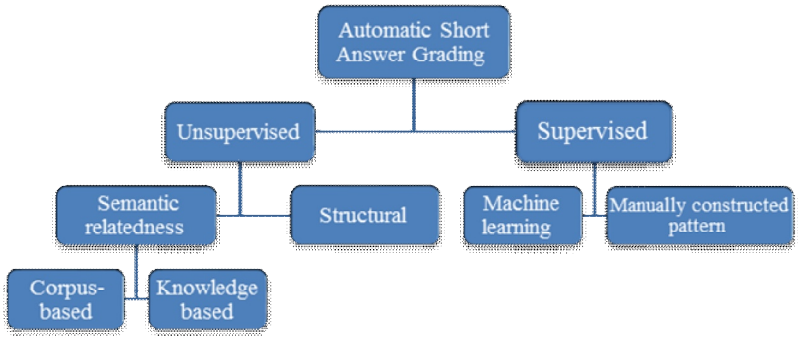

Figure 1: Classification of existing methods [16]

Unsupervised approaches in comparison to supervised methods do not need patterns, therefore they have higher flexibility and are confronted with less limitation [21-24].Unsupervised approaches can be classified into two groups of structural and semantic relatedness measures. Computing semantic relatedness among student's answer and the ideal answer is assumed as a highlighting technique in short answer grading systems. Semantic relatedness measures are also categorized into two types of knowledge-based and corpus-based[11, 25].

In corpus-based measures, the background knowledge is obtained by applying statistical analysis on the large collection of unsigned documents[13]. These methods generate a semantic space of terms where terms are distributed in the large corpus and terms' co-occurrences are used for computing semantic relatedness[26]. In these measures, each term is mapped to a multi-dimensional vector of concepts which represent implicit concepts[21-23]. On the other hand, knowledge-based measures leverage semantic relations among terms and concepts in background knowledge such as WordNet [27, 28]for computing semantic relatedness. In other words, these measures employ path length among two terms in a graph of concepts or semantic network of background knowledge for computing relatedness among terms[29].

3. Text Semantic Relatedness Measures

Whereas one of the purposes of this paper is to make comprehensive comparison among various text semantic relatedness and similarity measures in the application of short answer grading, in this section wide range of corpus-based and knowledge-based methods are studied. Since most of the knowledge-based measures are only able to compute relatedness among term, the methodology presented in [27]has been used for computing semantic relatedness among texts. Based on this methodology, for each term in the text, the maximum semantic relatedness score that can be achieved is considered. In other words, for each term W with part of speech C in an ideal answer, $maxsim(W, C)$ can be followed as follows:

$$maxsim(W, C) = maxSIM_x(W, w_i) \quad (1)$$

Where w_i is a term with part of speech of C in student answer and $SIM_x$is one the relatedness and similarity measures which are presented in the following section. The

semantic relatedness for each term is computed, summed together and normalized based on the length of both input texts.

3.1 Knowledge-based measures

In knowledge-based measures, the semantic relations of concepts defined in background knowledge are used to compute semantic relatedness. Some methods, especially earlier ones, have leveraged dictionaries and thesaurus[29]. Over time WordNet has changed into one of the most popular ontologies for computing semantic relatedness. The path was presented as a basic method which leveraged WordNet graph structure as ontology and considered inverse shortest path between two concepts [30]. The shorter the path from one node to another, the more related they are[30]:

$$rel_{path}(c_1, c_2) = \frac{1}{max\ len\ (c_1, c_2)} \quad (2)$$

This method performed fairly well but didn't take the graph depth into account. [31] normalized the path length using depth of the graph and solve Path's shortcoming:

$$rel_{Lch}(c_1, c_2) = -log\ \frac{len\ (c_1, c_2)}{2 \times \underset{c \in wordnet}{depth(\ c)}} \quad (3)$$

Where $len\ (c_1, c_2)$ shows the path length between two concepts and depth is the length of the longest path from the root node of the taxonomy to a leaf node[32].

Following the similar line of research, [32]proposed a measure which leveraged the notion of lowest common subsumer (LCS) of two concepts. LCS is the first shared concept from the leaf to the root of the hierarchy.

$$rel_{wup}(c_1, c_2) = \frac{2 \times depth\ (lcs)}{Depth\ (c_1) + depth(c_2\ )} \quad (4)$$

[33]also used WordNet graph structure for computing relatedness. Unlike above method which only took is-a relations between concepts into consideration, HSO used all of existing relations in WordNet.

$$rel_{HSO}(c_1, c_2) = C - len(c_1, c_2) - k.turns(c_1, c_2) \quad (5)$$

Where *C* and *K* are consonant, *len* is path length and turn is the frequency of direction changes between two concepts. Although the frequency of changes is less, semantic relatedness between two concepts is more.

Future more, some measures were presented which used the notion of information content. These approaches are based on this hypothesis that the relatedness of two concepts depends on the amount of information that they share. The first IC-based method is introduced by [34] which used WordNet as an ontology. Based on this measure the information content between two concepts was computed respect to their LCS. This means that if two pairs of terms have the same lowest common subsumer, the semantic relatedness between them will be equal:

$$rel_{res}(c_1, c_2) = IC\ \left(lcs(c_1, c_2)\right) \quad (6)$$

The information content of a concept is computed as:

$$IC(c) = -\log P(c) \quad (7)$$

Where p(c) is the probability of encountering an instance of a concept c in a large corpus. Resnik's definition of IC is widely used by later methods. Most of the later

To address Resink's problems,[35] proposed a measure. In this method, if a parent node is a subsumer, leaf nodes are used to compute semantic distance.

$$rel_{jcn}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2\,IC(lcs)} \quad (8)$$

On the other hand,[36] proposed a universal measure derived from information content. In the beginning, it was only applied to taxonomic structures.

$$rel_{lin}(c_1, c_2) = 2.\frac{IC(lcs)}{IC(c_1) + IC(c_2)} \quad (9)$$

In addition, WordNet contains a small definition for each term which clarifies the meaning of the corresponding term. Lesk proposed a semantic relatedness measure which is based on the definition of terms in Wordnet known as a gloss. In other words, the Lesk measure assigns relatedness by finding and scoring overlaps between the glosses of the two concepts[37]. Vector is another knowledge-based measure which creates a co-occurrence matrix from a corpus made up of the WordNet glosses. Based on this measure, a vector is generated for each term and they are compared with each other for computing relatedness[27].

3.2 Corpus-based measures

Corpus-based models leverage statistical analysis on background corpus to build semantic space. In other words, unlike knowledge-based methods, these methods do not require explicit relations among concepts and are able to compute relatedness between terms based on their co-occurrences in a large corpus of documents. Based on this theory, terms that co-occur in the same context tend to be related[29].

Going beyond simple co-occurrence, Latent semantic analysis (LSA)[38] is one of the most important measures which uses vector presentation for computing relatedness and it is able to discover hidden structure among terms. It is a dimensional reduction approach, which applies Singular Value Decomposition (SVD) on a term-document matrix in order to map terms to latent topics and generalize observed relations between terms and topics[38].

Another approach was introduced by [39] and it was referred as Explicit Semantic Analysis (ESA). Based on ESA, vectors constructed from Wikipedia concepts are used for computing relatedness. In other words, ESA relies exclusively on distributional similarity mechanism. Moreover, because of employing Wikipedia articles, which are understandable to humans, this model is explicit and presents high correlation coefficient with human judgments[39].

It must be taken into consideration that in spite of existing large number of measures for computing semantic relatedness and similarity of texts, a comprehensive study considering performance comparison of these measures in the application of short answer grading has not been accomplished yet. Therefore, the first goal of this paper is to clarify the performance of these measures in the field of short answer grading. In the following, it is supposed to present an approach that can be able to increase the precision of these measures in is this field.

4. An Approach Based on Automatic Pseudo-Relevance Feedback

Automatic short answer grading using semantic relatedness measures can be done by comparing student's answer and an ideal answer. Whereas in exams based on the text the focus is on the concept of given answers and there is only one ideal answer, an answer despite having correct concept can be assigned by the low score.

To fill this lacuna, a novel approach based on the feedback of students' answers with the highest score is proposed which is similar to the pseudo feedback method in information retrieval[40,42]. Using this technique and interpreting students' answers, the number of words in the ideal answer can be increased. In other words, correct answers can be used to expand ideal answer and in the following, the precision of semantic relatedness measures in the application of automatic short answer grading will be increased. The general overview of the proposed approach is presented in figure 2.

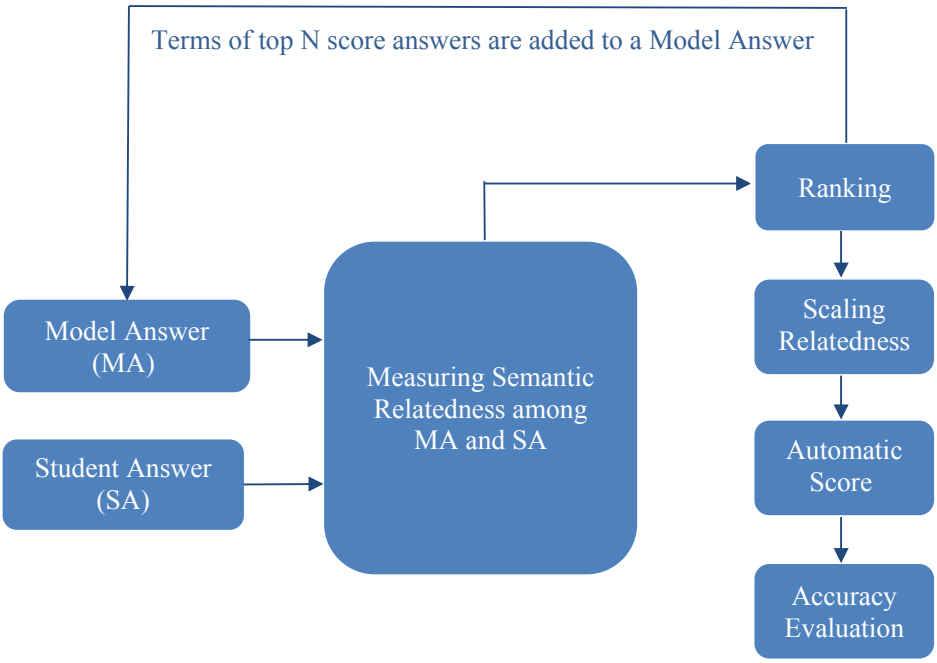

Figure 2: Overview of the Proposed Approach

As it is shown, Firstly, the semantic relatedness among students' answers and the ideal answer is computed using various measures presented in section 3. Secondly, assigned scores are sorted in decreasing order and then terms of N answers with the highest score are added to ideal answer (the ideal answer is expanded using terms of students' answers with the highest score). This step is inspired by similar to pseudo-relevance feedback used in information retrieval. Thirdly, remaining answers are scored again using a new ideal answer obtained from the last step. Indeed, N answers with the highest score are not graded again (using feedback), but the other answers are graded again using the new expanded ideal answer. Based on this strategy, the highest score of first N answer is saved and it is guaranteed that none of the remaining answers have a higher score than them.

While the semantic scores gained from semantic relatedness measures are between zero and one and semantic scores of the dataset used in experiments is between one and

five, scores achieved from semantic relatedness algorithms must be scaled in order to assign scores to the answer. Finally, in order to specify the performance of the proposed method, the Pearson Correlation Coefficient between scores achieved from the algorithms and human grading (gold standard) is computed. Results of experiments, presented in next section, revealed that using pseudo-feedback can considerably increase the precision of semantic relatedness measures in the application of short answer grading. Notably, maximum improvement observed after about 4-6 iterations on average.

5. Implementation and Empirical Results

Experiments can be divided into two groups. In the first set of experiments, corpus-based and knowledge-based semantic relatedness measures are compared with each other in the application of short answer grading. In the second set of experiments, the effect of using pseudo-relevance feedback on the performance of these measures is taken into consideration.

5.1 Dataset

Results of experiments are evaluated using mathematical analysis and computing the correlation between empirical results and human judgment. Whereas in the field of automatic short answer grading there are datasets containing sets of questions and students' answers which are scored by some teachers, various semantic relatedness measures can be applied to these datasets and by computing the Pearson correlation coefficient among empirical results the precision of these measures in application of short answer grading can be revealed[11].

Dataset employed in this paper contains tree exercise that each of them includes seven questions and short answers which are answered by 30 students. Therefore, this dataset contains 630 answers. Answers are individually scored by two annonator in the range of zero (completely wrong) to five (completely correct) and The correlation between the two human judges is measured (r=0.7228)[23]. Notably, it is a standard dataset in the field of ASAG and large number of researches have been applied on it. Table 1 shows two question answer pairs with three sample student answers each are presented in Table1. The assigned grades by the two human judges are also included.

Table 1-Two sample questions with short answers provided by students and the grades assigned by the two human judges[23]

Sample questions, correct answers, and student answers	Grades	
Question: What is the role of a prototype program in problem solving? Correct answer: To simulate the behavior of portions of the desired software product.		
Student answer 1: A prototype program is used in problem solving to collect data for the problem.	1	2
Student answer 2: It simulates the behavior of portions of the desired software product.	5	5
Student answer 3: To find problem and errors in a program before it is finalized.	2	2
Question: What are the main advantages associated with object-oriented programming? Correct answer: Abstraction and reusability.		
Student answer 1: They make it easier to reuse and adapt previously written code and they separate complex programs into smaller, easier to understand classes.	5	4
Student answer 2: Object oriented programming allows programmers to use an object with classes that can be changed and manipulated while not affecting the entire object at once.	1	1
Student answer 3: Reusable components, Extensibility, Maintainability, it reduces large problems into smaller more manageable problems	4	4

5.2 Implementation

In order to evaluate various measures of semantic relatedness in the application of short answer grading comprehensively, measures presented in section 2 are implemented. In the following implementation details of each measure is expressed respectively. It must be noted that although different configurations and implementation are available for each measure, it has been tried to choose the best one according to the previous studies.

WordNet::Similarity package is employed for implementing measures which used WordNet as background knowledge[30]. Gensim package is also used for implementing LSA[41]. For implementing ESA, the methodology presented by[39] is used. Noteworthy, the results obtained by each measure are normalized to be in the range of zero to one. Moreover, WordNet 1.2 and Wikipedia 2015 has been employed as background knowledge for implementing various knowledge-based and corpus-based semantic relatedness measures.

5.3 Experimental Results

In this section, comprehensive experiments have been done to evaluate the performance of semantic relatedness measures in the application of short answer grading. The experiments of this paper are divided into 3 parts. In the first part of experiments, the performance of corpus-based and knowledge based semantic relatedness measures in the task of automatic short answer grading is explored. While the performance of corpus-based methods is highly dependent on background knowledge, the effect of domain and size of background knowledge is investigated in the second part of experiments. The effect of using the proposed method on the precision of semantic relatedness measures in the application of short answer grading is consider in the third part of experiments.

According to the first part of experiments, the correlation among results achieved by experiments and human judgments using Pearson correlation coefficient is shown in table 2. Pearson correlation among scores obtained from algorithms and human scores is used to clarify the precision of each algorithm. In other words, Pearson correlation is used to specify how results obtained from algorithms are similar to human judgments.

Results of experiments revealed that the performance of corpus-based measures in significantly better that knowledge-based measures in the application of short answer grading. Considering that the average correlation score between corpus-based measures with human judgment ($\rho^*_{corpus} = 0.475$) is significantly higher than the average correlation score between knowledge-based models with human judgment ($\rho^*_{knowledge} = 0.332$). The best knowledge-based model has the correlation of $\rho_{path} = 0.451$ with human judgments, while the best corpus-based model has the correlation of $\rho_{ESA} = 0.498$. The lowest correlation refers to HSO [33]and the highest one refers to ESA[39].

It must be taken into consideration that knowledge-based measures require structured background knowledge that its construction is very costly and time-consuming. In contrast, corpus-based measures use unstructured corpora which are confronted with less limitations in real world problems.

Table 2: Comparison between corpus-based and knowledge-based measures of semantic relatedness in the application of short answer grading

Type	Algorithm	Pearson Correlation Coefficient
Knowledge-based measures	Path [30]	0.451
	LCH [31]	0.223
	Lesk [37]	0.363
	WuP [32]	0.336
	Resnik [34]	0.252
	Lin [36]	0.391
	JCN [35]	0.449
	HSO [33]	0.196
	Vector [27]	0.382
Corpus-based measures	ESA [39]	**0.513**
	LSA[38]	0.438

As previously mentioned, the second part of experiments refers to investigating the domain and size of exploited background knowledge. In other words, the performance of corpus based measures, despite of their superior performance, is related to the used background knowledge and they are expected to be sensitive to its size and domain. In this regard, three training corpora were used in our experiments as Wikipedia full, Wikipedia small and Wikipedia specific. Wikipedia full is the open domain corpus of Wikipedia containing all articles which was used in the last part of experiments. Wikipedia small is random subset of Wikipedia with smaller size. Wikipedia specific is also a subset of Wikipedia containing articles in the field of computer science. The results of employing various kind of corpora along with the Pearson correlation is presented in table 3.

Table 3: effect of various domain and size of background knowledge on the performance of corpus based measures

Algorithm	Size	Pearson Correlation Coefficient
LSA Wikipedia	2.1 GB	0.438
LSA Small	4 MB	0.357
LSA Specific	78 MB	0.468
ESA Wikipedia	2.1 GB	0.513
ESA Small	4 MB	0.414
ESA Specific	78 MB	0.483

By comparing the results of applying LSA in various corpora, it is observed that by employing a domain specific corpus, higher correlation is obtained. It can be concluded that the quality of text generally more important that the quantity of text for LSA. In contrast, by employing domain specific subset of Wikipedia, lower correlation is achieved. Therefore, it can be stated that for ESA the high dimensionality of concept space is very crucial.

The third set of experiments express the effect of the proposed approach in section 4 on the precision of semantic relatedness measures in the application of short answer grading. The result obtained from using pseudo-feedback with the aim of improving the precision of semantic relatedness measures in the application of short answer grading is presented in table 4. The results of experiments have shown that using pseudo-feedback can improve the precision of all semantic relatedness measures (Path algorithm is used as representative of knowledge-based measures). As it is clear, using Pseudo-feedback

can significantly improve the precision of LSA and Path methods in comparison to ESA.

Table 4: Experimental results after using pseudo-feedback

Algorithm	Pearson Correlation Coefficient
Path	0.488
ESA	0.523
LSA	**0.545**

Figure 3 presents the impact of automatic feedback based on the size of N. This diagram is drawn based on the number of answers with high scores used in feedback and Pearson correlation coefficient. As it is clear, increasing the number of N (more than 10) can cause a reduction in precision of algorithms. The diagram illustrates that LSA measure is more sensitive to automatic feedback than other algorithms and using automatic feedback can have a significant impact on the performance of this measure. Accordingly, this measure can extensively be employed in the application of short answer grading.

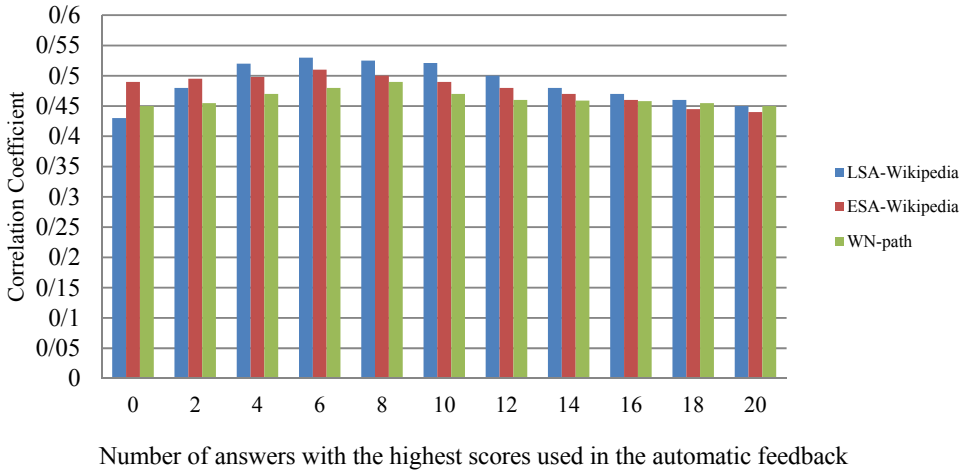

Figure 3: The effect of using automatic feedback on semantic relatedness measures

6. Conclusion

- In this paper, various measures of semantic relatedness are compared to each other in the application of short answer grading. For this aim, existing measures are divided into two groups of knowledge-based and corpus-based measures and extensive experiments are carried out to clarify the performance of these measures in the field of short answer grading. The results of empirical experiments present that corpus-based measures have higher precision in comparison to knowledge-based measures in the application of short answer grading and they are also confronted with fewer limitations in real world problems. In other words, knowledge-based measures require structured background knowledge that needs the human resource for construction. Notably, constructing this background knowledge is also costly and time-consuming.

- In the following, a novel method was presented which combines students' answers with an ideal answer for improving the precision of semantic relatedness measures in the application of short answer grading. The presented method performs same as pseudo-feedback in information retrieval and expands the ideal answer using students' answers with the highest scores which cause significant improvement in precision of semantic relatedness measures in the application of short answer grading. Empirical results present that using automatic feedback has the considerable effect on LSA measure and can increase its precision up to 0.53.
- For future work, with the aim of increasing the precision of semantic related measures in the application of short answer grading, they can be combined with machine learning algorithms. Deep learning methods can also be used for this aim. Moreover, by constructing a domain-specific background knowledge, it can be possible to design a system which can grade short answers in other languages such as Persian.

References

[1] Smith, P.A.M., et al., *A Multimodal Assessment Framework for Integrating Student Writing and Drawing in Elementary Science Learning*. IEEE Transactions on Learning Technologies, 2018.

[2] Pribadi, F.S., et al. *Automatic short answer scoring using words overlapping methods*. in *AIP Conference Proceedings*. 2017. AIP Publishing.

[3] Burrows, S., I. Gurevych, and B. Stein, *The Eras and Trends of Automatic Short Answer Grading*. International Journal of Artificial Intelligence in Education, 2015. **25**(1): p. 60-117.

[4] Shermis, M.D. and J. Burstein, *Handbook of automated essay evaluation: Current applications and new directions*. 2013: Routledge.

[5] Polson, M.C. and J.J. Richardson, *Foundations of intelligent tutoring systems*. 2013: Psychology Press.

[6] Kohail, S .and C. Biemann. *Matching, Reranking and Scoring: Learning Textual Similarity by Incorporating Dependency Graph Alignment and Coverage Features*. in *18th International Conference on Computational Linguistics and Intelligent Text Processing. Budapest, Hungary.* 2017.

[7] Mousavi, H., *Automatic Short Essay Scoring Using Natural Language Processing to Extract Semantic Information in the Form of Propositions*. 2013, University of California, Los Angeles.

[8] Amini, B., R. Ibrahim, and M.S. Othman, *Adaptive Information Analysis in Higher Education Institutes*. Journal of Advances in Computer Research, 2011. **8**(4): p. 1-12.

[9] Bahrami, N., A.H. Jadidinejad, and M. Nazari, *Computing Semantic Similarity of Documents Based on Semantic Tensors*. Journal of Information System and Telecommunication. 2015. **3** (2): p- 125-134.

[10] Zhang, Y., R. Shah, and M. Chi. *Deep Learning+ Student Modeling+ Clustering :a Recipe for Effective Automatic Short Answer Grading*. in *EDM*. 2016.

[11] Zhang, Z., A.L. Gentile, and F. Ciravegna, *Recent advances in methods of lexical semantic relatedness–a survey*. Natural Language Engineering, 2013. **19**(4): p. 411-479.

[12] Young, J.R., *Inside the Coursera contract: How an upstart company might profit from free courses*. The Chronicle of Higher Education, 2012. **19**(07): p. 2012.

[13] Butcher, P.G. and S.E. Jordan, *A comparison of human and computer marking of short free-text student responses*. Computers & Education, 2010. **55**(2): p. 489-499.

[14] Allahverdipour, A. and F. Soleimanian Gharehchopogh, *A New Hybrid Model of K-Means and Naïve Bayes Algorithms for Feature Selection in Text Documents Categorization*. Journal of Advances in Computer Research, 2017. **8**(4): p. 73-86.

[15] Bostan, S. and M. Ghasemzadeh, *Personalization of Search Engines, Based-on Comparative Analysis of User Behavior*. Journal of Advances in Computer Research, 2015. **6**(2): p. 65-72.

[16] Jadidinejad, A.H. and F. Mahmoudi ,*Unsupervised Short Answer Grading Using Spreading Activation over an Associative Network of Concepts*.Canadian Journal of Information and Library Science, 2014. **38**(4): p. 287-303.

[17] Nazari Soleimandarabi, M., S.A. Mirroshandel, and H. Sadr, *A Survey of Semantic Relatedness Measures*. International Journal of Computer Science and Network Solutions, 2015. **3** (2): 12-23.

[18] Nazari Soleimandarabi, M., S.A. Mirroshandel, and H. Sadr, *The Significance of Semantic Relatedness and Similarity measures in Geographic Information Science*. International Journal of Computer Science and Network Solutions, 2015. **3** (2): 12-23.

[19] Jadidinejad, A.H. and H. Sadr, *Improving weak queries using local cluster analysis as a preliminary framework*. Indian Journal of Science and Technology, 2015. **8**(15).

[20] Nazari Soleimandarabi, M. and S.A. Mirroshandel, *A novel approach for computing semantic relatedness of geographic terms*. Indian Journal of Science and Technology, 2015. **8**(27).

[21] Leacock, C. and M. Chodorow, *C-rater: Automated Scoring of Short-Answer Questions*. Computers and the Humanities, 2003. **37**(4): p. 389-405.

[22] Mohler, M., R. Bunescu, and 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 - Volume 1*. 2011, Association for Computational Linguistics: Portland, Oregon. p. 752-762.

[23] Mohler, M. and R. Mihalcea. *Text-to-Text Semantic Similarity for Automatic Short Answer Grading*. in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. 2009. Athens, Greece: Association for Computational Linguistics.

[24] Ziai, R., N. Ott, and D. Meurers. *Short answer assessment: Establishing links between research strands*. in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. 2012. Association for Computational Linguistics.

[25] Sadr, H., M. Nazari Solimandarabi, and M. Mirhosseini Moghadam, *Categorization of Persian Detached Handwritten Letters Using Intelligent Combinations of Classifiers*. Journal of Advances in Computer Research, 2017. **8**(4): p. 13-21.

[26] Baroni, M. and A. Lenci, *Distributional Memory: A General Framework for Corpus-Based Semantics*. Computational Linguistics, 2010. **36**(4): p-721-673

[27] Patwardhan, S. and T. Pedersen. *Using WordNet-based context vectors to estimate the semantic relatedness of concepts*. in *EACL Workshop Making Sense of Sense* - Workshop of Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. 2006.

[28] Jarmasz, M. and S. Szpakowicz, *Roget's Thesaurus and Semantic Similarity*. CoRR, 2012. abs/1204.0245.

[29] Zesch, T. and I. Gurevych, *Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words*. Natural Language Engineering, 2010. **16**(1): p. 25-59.

[30] Pedersen, T., S. Patwardhan, and J. Michelizzi. *WordNet::Similarity: Measuring the Relatedness of Concepts*. in *Demonstration Papers at HLT-NAACL 2004*. 2004. Stroudsburg, PA, USA: Association for Computational Linguistics.

[31] Leacock, C. and M. Chodorow, *Combining local context and WordNet similarity for word sense identification*. WordNet: An electronic lexical database, 1998. **49**(2) :p. 265-283.

[32] Wu, Z. and M. Palmer. *Verbs semantics and lexical selection*. in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. 1994. Association for Computational Linguistics.

[33] Hirst, G. and D. St-Onge, *Lexical chains as representations of context for the detection and correction of malapropisms*. WordNet: An electronic lexical database, 1998. **305**: p. 305-332.

[34] Resnik, P., *Using information content to evaluate semantic similarity in a taxonomy*. arXiv preprint cmp-lg/9511007, 1995.

[35] Jiang, J.J. and D.W. Conrath, *Semantic similarity based on corpus statistics and lexical taxonomy*. arXiv preprint cmp-lg/9709008, 1997.

[36] Lin, D., *An Information-Theoretic Definition of Similarity*, in *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998, Morgan Kaufmann Publishers Inc. p. 296-304.

[37] Lesk, M., *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*, in *Proceedings of the 5th annual international conference on Systems documentation*. 1986, ACM: Toronto, Ontario, Canada. p. 24-26.

[38] Dumais, S.T., *Latent semantic analysis*. Annual Review of Information Science and Technology, 2004. **38**(1): p. 188-230.

[39] Gabrilovich, E. and S. Markovitch, *Wikipedia-based semantic interpretation for natural language processing*. J. Artif. Int. Res., 2009. **34**(1): p. 443-498.

[40] Baeza-Yates, R. and B. Ribeiro-Neto, *Modern information retrieval*. Vol. 463. 1999: ACM press New York.

[41] Řehůřek, R. and P. Sojka. *Software Framework for Topic Modelling with Large Corpora*. in *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. 2010. Valletta, Malta: University of Malta.

[42] Sadr, H., RE. Atani, MR. Yamaghani, *The Significance of Normalization Factor of Documents to Enhance the Quality of Search in Information Retrieval Systems*. International Journal of Computer Science and Network Solutions.2014. **2** (5), 91-97.