*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2634–2643
Marseille, 11–16 May 2020


# AR-ASAG An ARabic Dataset for Automatic Short Answer Grading Evaluation

**Leila OUAHRANI, Djamal BENNOUAR**
Bouira University
Algeria
l_ouahrani@univ-blida.dz, djamal.bennouar@univ-bouira.dz

**Abstract**

Automatic short answer grading is a significant problem in E-assessment. Several models have been proposed to deal with it. Evaluation and comparison of such solutions need the availability of Datasets with manual examples. In this paper, we introduce AR-ASAG, an Arabic Dataset for automatic short answer grading. The Dataset contains 2133 pairs of (Model Answer, Student Answer) in several versions (txt, xml, Moodle xml and .db). We explore then an unsupervised corpus based approach for automatic grading adapted to the Arabic Language. We use COALS (Correlated Occurrence Analogue to Lexical Semantic) algorithm to create semantic space for word distribution. The summation vector model is combined to term weighting and common words to achieve similarity between a teacher model answer and a student answer. The approach is particularly suitable for languages with scarce resources such as Arabic language where robust specific resources are not yet available. A set of experiments were conducted to analyze the effect of domain specificity, semantic space dimension and stemming techniques on the effectiveness of the grading model. The proposed approach gives promising results for Arabic language. The reported results may serve as baseline for future research work evaluation.

**Keywords:** Short Answer Grading, Dataset, Semantic Space, Semantic Similarity, Term Weighting, Arabic Language, Corpus, Stemming

## 1. Introduction

Assessment is a key component of the teaching and learning process. The templates provided by most Computer-Aided Assessment Systems are multiple-choice questions, true/false questions and matching questions. Only basic support around the management of open-ended questions (short answer questions and essays) is offered by few others. Short answer questions(few words to a few sentences constructed in natural language) are more effective focusing on recall and reproduction (Anderson et al., 2001). However, their assessment is a complex and subjective process that requires the analysis and a deep understanding of a natural language text. Although Automatic Short Answer Grading systems (ASAG) have been studied for many years, their adoption in practice is not common due to their complexity (Liu et al., 2014). Unlike some systems which use information extraction with manually written patterns, templates, or machine learning to perform the ASAG task, we deal with the problem as a semantic similarity problem between the Student's Answer (SA) and the teacher's Model Answer (MA)(( Mohler and Mihalcea, 2009), (Mohler et al., 2011), (Gomaa and Fahmy, 2014b), (Gomaa and Fahmy, 2014a), (Zahran et al., 2015), (Magooda et al., 2016), (Bennouar, 2017)). There are two main approaches to determine the semantic similarity between two short texts: topological similarity and statistical similarity(corpus-based) (Mihalcea et al., 2006). Topological similarity uses data models containing information about concepts and their correlation (WordNet[1], thesaurus, dictionaries). Statistical similarity uses vector state spaces to express correlations of words extracted from text corpora. Many solutions for calculating semantic similarity, both topological and statistical, have already been developed for English. However, few are designed that they can be adapted to under-resourced languages like Arabic language, because they often use advanced natural language processing techniques which are specific to a language. Researchers have developed a wide range of NLP tools to analyze, parse and annotate different languages automatically. Language resources play two roles in these activities. The first is the use of large-scale annotated corpora to drive statistical NLP techniques. The second is the need for test collections(Datasets) for the purpose of evaluation against a gold-standard. Such resources for NLP are documented by efforts such as the Language Resources and Evaluation Map(Gratta et al., 2014). But for some languages, there are few such resources. Arabic is an appropriate example to consider. Despite being a widely spoken language, it has been widely acknowledged that it has few publicly available tools and resources, apart from a few notable exceptions(Mahmoud El-Haj et al., 2015). In particular, Arabic NLP lacks resources such as corpora, lexicons, machine-readable dictionaries, Datasets in addition to fully automated fundamental NLP tools such as tokenizers, part-of-speech taggers, parsers, stemmers and semantic role labelers. A lack of sufficient data and research has negatively affected Arabic natural language processing practitioners(Mahmoud El-Haj et al., 2015). Arabic WordNet(AWN[2]) developed with the same methodology as WordNet lacks a lot of information and concepts and semantic relationships between synonym-sets. In the other hand, most datasets cannot be shared for reasons such as privacy. Frequently, academics are simply adapting data from their own teaching experiences to ASAG projects, but with little consideration that others may want to perform meaningful comparisons to their methodology. Authors dealing with the Arabic short answer grading evaluate their models on punctual examples. No Arabic Dataset is publicly available. To deal with this double issue, we introduce, in this paper, AR-ASAG; an Arabic Dataset for Automatic Short Answer Grading and we explore an unsupervised vector space model for automatic grading adapted to Arabic Language to face the challenge of resources lack. The only resources requirements language-dependent are an undifferentiated text corpus and a stemmer. The grading process is based on COALS (Correlated Occurrence Analogue to Lexical Semantic)(Rohde et al., 2004) algorithm that gives distributional word representation based on co-occurrences in text corpora. The summation vector model is combined

[1] http://globalwordnet.org/

[2] http://globalwordnet.org/arabic-wordnet/

to term weighting and common words to achieve similarity between a Model Answer and a Student Answer. The developed Dataset is used in experiments to specifically seek answers to the following research questions.

*First*, using a semantic space approach, to what extent does the domain and dimension of semantic influence the accuracy of the grading?

*Second*, how can word weighting improve the quality of grades for a grading system that is easy to implement in practice? Since very few work in Arabic language use word weighting?

*Finally*, what effect of stemming techniques on grading accuracy for a language as inflectional as Arabic?

The reported experimental results may serve as a baseline for other researchers interested in automatic short answer grading evaluation.

## 2. Related Work

Various approaches have been proposed for automatic grading of short answers. (Burrows, Gurevych, and Stein, 2015) and (Shourya Roy, Y. Narahari, 2015) gave a comprehensive review of ASAG systems. Here we briefly discuss closely related work dealing with automatic grading of short answers as a text to text similarity task and using vector word distribution. The Text-to-Text Texas system in English was introduced in (Mohler and Mihalcea, 2009). The score is assigned according to a measure of the semantic similarity between a student answer and a model answer. Several measures, including knowledge-based and corpus-based are used in this approach. The system was applied to a computer science dataset (Texas dataset[3]) that contains 21 questions and 630 student responses. Student answers are scored in the interval [0...5] by two human expert annotators with a Pearson correlation r=0.6443. The best Pearson correlation value between the automatic and manuals scores was 0.47. This system was enhanced in (Mohler et al., 2011). Few research studies dealing with automatic grading of Arabic short answer have been published. We focus here on works that used datasets and presented evaluation results ((Gomaa and Fahmy, 2014b), (Magooda et al., 2016), (Zahran et al., 2015)). In (Gomaa and Fahmy, 2014b), before applying multiple similarity measures separately and in combination, authors translated students' answers into English to overcome the lack of text processing resources in Arabic language. Additionally, this research(Gomaa and Fahmy, 2014b) presented the first (and the only one to our knowledge) Arabic dataset (the Cairo university dataset). The dataset is not publicly available *(Authors accepted to share an XML version with us in a precedent work).* In Cairo University Arabic Dataset(Gomaa and Fahmy, 2014b), questions cover a chapter of the official Egyptian curriculum for the Environmental Science course. The dataset contains 61 questions, 10 answers for each, with a total of 610 answers with their English translations. Student answers are scored in the interval [0...5] by two human expert annotators with a Pearson correlation r=0.86 and a RMSE =0.69. In (Zahran et al., 2015), authors compare different techniques to build vectorized space representations for Arabic language. In (Magooda et al., 2016) the research exploited several vector representations to various sentence representations techniques. A wide range of similarity measures are compared and finally a system is proposed combining nine measures using Zahran vector representations(Zahran et al., 2015). As short Answer Grading and text similarity tasks are strongly related in our proposed approach, we consider SEMEval-2017(Semantic Textual Similarity-Multilingual and Cross-lingual Focused Evaluation)(Agirre et al., 2017) Workshop which makes available STS 250 SEMEval 2017 Dataset[4]for track 1 Arabic-Arabic. The Dataset contains 250 pairs of sentences obtained by translation from English into Arabic. For each pair, a manual gold score that averages five human annotations, is given. Authors in (Nagoudi and Ferrero, 2017) proposed a system LIM-LIG (the second score in track 1). They used Zahran Word Embedding (Zahran et al., 2015). STS 250 SEMEval 2017 Dataset is used in this paper to evaluate our grading model and to make achievements comparison with the developed AR-ASAG Dataset.

## 3. The AR-ASAG Dataset

To support the evaluation of short answer grading solutions, we created a new Dataset for Arabic Language. To our knowledge, such Dataset represents the first Arabic Dataset *made publicly available*. The Dataset will be of value as a resource evaluation for other researchers working on short answer grading and semantic similarity in Arabic language. The Dataset consists of questions extracted from the teaching course on cybercrimes with answers provided by three classes of master students. To have a Dataset representative of the reality it was necessary to pass by the teaching of the course in Arabic language. About 170 students having *native Arabic language* followed the course which was validated by an official exam. We make our dataset freely available (AR-ASAG Dataset, 2020).

### 3.1 Data collection

The reported evaluations relate to answers submitted for three different exams submitted to three classes of students. The exams were conducted under natural conditions of evaluation. Each test consists of 16 short answer questions (a total of 48 questions). The Dataset encompasses 5 types of questions:

- "عرف ": *Define?*
- "إشرح": *Explain?*
- "ما النتائج المترتبة على": *What consequences?*
- "علل": *Justify?*
- "ما الفرق": *What is the difference?*

Students submitted answers to the questions. The number of answers obtained is different from one question to another. Identical student answers are only reported once in the dataset. Thus, our dataset includes a total of 2133 student answers. For each question, a model answer is proposed. In Fig. 1, we present the distribution of answers per type of question. Question types are identified respectively 1, 2, 3, 4 and 5.

[3] http://web.eecs.umich.edu/~mihalcea/downloads.html#saga

[4]http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

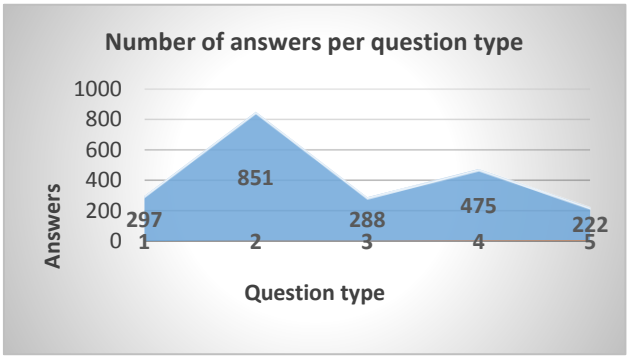


Figure 1: Number of answers per question type

The answers were independently graded by two human experts, using a scale from 0 (completely incorrect) to 5 (perfect answer). Both human experts were computer science teachers. We treat the average grade of the two annotators as the gold standard against which we compare our proposed system outputs. Table 1 shows a question-answer pair with three sample student answers and the two manual grades assigned by experts.

| **Sample Questions, Reference Answers, Student Answers and the two Manual Grades (AR-ASAG Dataset)** | | | |
|---|---|---|---|
| ***Question*** | عرف مصطلح الجريمة المنظمة على الانترنت<br>*Define: Online Organized Crime* | | |
| ***Reference answer*** | عنف منظم تقوم به جماعات ترتكب أفعالا تخترق بها القانون للحصول على مكاسب مالية، بطرق وأساليب غير مشروعة تنفذ بعد تدبير وتنظيم<br>*It is an organized violence by groups committing acts to gain financial gain, in unlawful ways using measure and organization.* | | |
| ***Student answer 1*** | هي عنف منظم تقوم به جماعات من اجل كسب الأموال وتعتمد على التنظيم وكسب الأموال.<br>*It is an organized violence organized by groups to make money and depends on organization and making money.* | 5 | 5 |
| ***Student answer 2*** | هي سلوك غير قانوني تقوم على التنظيم بهدف سرقة المعلومات او تغييرها.<br>*It is an illegal behavior based on the purpose of stealing or changing information.* | 4.5 | 4 |
| ***Student answer 3*** | عنف منظم يسعى من خلاله تحقيق مطالب مالية غير شرعية تقع على الأنترنت<br>*Organized violence which seeks to achieve illegal financial requests using Internet* | 2.5 | 3 |

Table 1: Sample question, Reference Answer, Student Answers and the two Manual Grades

## 3.2 Inter-Annotator Agreement

Assigning manual grades was a challenging task for the annotators. They noted that the difficulty lay not in deciding whether or not two answers were semantically similar, but in determining the precise degree of similarity and then the grade. To understand how consistent, the annotators were with one another, evaluations are run using Pearson's correlation coefficient (r: the higher the better)) and the Root Mean Squared Error (RMSE: the lower the better) measured against the average of the human-assigned grades on a per-question basis. Every question and the corresponding student answer is considered as an independent data point, and thus the emphasis is placed on the correctness of the grade assigned to each answer. The correlation between the two experts and the Error are measured. The two annotators correlated at *(r=0.8384)* with an *(RMSE=0.8381)*. Automated Short Answer Assessment is a subjective assessment that emphasizes on contents. Since subjectivity is consubstantial with any evaluative act(Brown et al., 1999), a closer examination of the grades assigned by the two human annotators indicates the underlying subjectivity in the grading of short-answer questions. Indeed, as shown in Table 2, in 34.83% (743 answers) both annotators gave the same grades. In 54.14%(1155 answers), the difference is at most one point. However, in 11.01%(235 answers), the difference is more than one point. In 2.15% (46 answers) the difference is more than 2 points on a scale of five points.

| **Difference** | ***Number of answers*** | ***%*** |
|---|---|---|
| 0 | 743 | 34.83 |
| 0 <D <= 1 | 1155 | **54.14** |
| 1 <D <= 2 | 189 | 8.86 |
| 2 <D <= 3 | 38 | 1.78 |
| D > 3 | 8 | 0.37 |

Table 2: Annotator Analysis

Moreover, when the two annotators disagreed, the second annotator gave the higher grade 38.2% of the time. The average grade given by grader1 is 2.86, while the average grade given by grader2 is 2.94 for the complete Dataset. This subjectivity is apparent in the distribution of deviation grades between the two annotators illustrated in Fig. 2. In addition to the RMSE error for the Dataset, we report the median error RMSE for each question. Deviation is considered for each answer, the Av(RMSE) = 0.5629. This gives an indication of the Inter-Annotator Agreement allowing a single question to be noted in isolation.

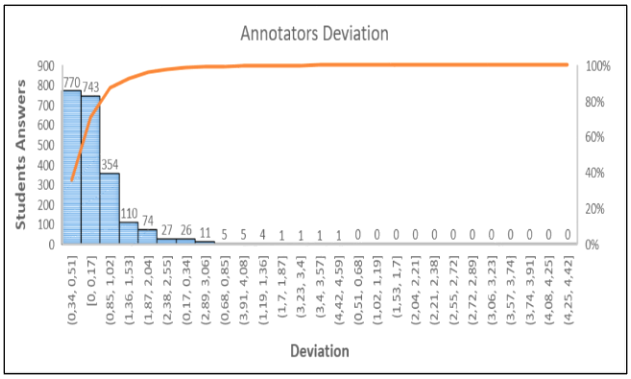


Figure 2: Inter-Annotator Agreement

Subjectivity can be explained by the diversity of evaluation criteria from one annotator to another, guided by different frameworks to judge student answers even if the model answers already exist. Note that annotators have not received explicit instructions on how to assign grades other than the [0..5] scale.

## 3.3 Question Demoting

We proceeded to a question demoting of the Dataset to avoid rewarding the response of a student who repeats question words. We removed the words in the text of the

question from the reference answer and the student's answers.

### 3.4 Dataset Versions

AR-ASAG Dataset is available in different versions: TXT, XML, XML-MOODLE and Database (.DB). The .DB format allows to make the necessary exports according to specific analysis needs. This format enables easy and efficient evolution of the Dataset. The dataset can continuously grow by introducing new tests and exams into the database. All exports to the different versions are done automatically by the dataset management program. The XML-MOODLE format is especially interesting when it is considered for the evaluation of short answers grading systems in the context of the MOODLE[5] platform (widely used as e-learning platform). The Dataset can be used as a questions bank and also to compare performance with the short answer system developed on the Moodle platform based on grammars and patterns matching[6]. Note here that both manual grades are available in the dataset. This allows a thorough analysis of the behavior of the automatic system with that of human annotators, especially for such a subjective domain where no agreement on evaluation criteria exists yet. In Table 3, AR-ASAG dataset is compared to the frequently used datasets in evaluation of automatic assessment.

| *Dataset* | *Lang.* | *Answers* | *Domain* | *Availability* | *IAA(Pearson)* |
|---|---|---|---|---|---|
| Texas Dataset *(Mohler and Mihalcea, 2009)* | English | 630 | Data Structures Course | yes | 0.6443 |
| Cairo University Dataset *(Gomaa and Fahmy, 2014b)* | Arabic | 610 | Environmental Science Course | no | 0.8600 |
| AR-ASAG Dataset | Arabic | 2133 | Cybercrimes Course | yes | 0.8384 |

Table 3: AR-ASAG Dataset vs. ASAG Datasets

## 4. Automatic Short Answer Grading

For automatic short answer grading, our experiments focus on the use of semantic space based similarity measure combining term weighting and index of common words between the model answer and the student answers. The reported result of experiments may serve as a baseline for future researchers using AR-ASAG Dataset. In particular, we conducted a set of experiments, seeking for answers to the following research questions dealing with Arabic.

*First*, using a semantic space approach, to what extent does the domain and dimension of semantic space influence the accuracy of the grading model? To answer this question, we conduct a series of experiments with different dimensions and domains of the corpora used to construct semantic spaces. We then measured their effect on the quality of short answer grading. Here we used three available Arabic corpora and create our own specific corpus domain.

*Second*, how can word weighting improve the quality of grades for an Arabic grading system? To answer this question, we consider the NTFlog (Normalized TFlog) weighting in addition to IDF (Inverse Document Frequency).

*Finally*, what is the effect of stemming on grading accuracy for Arabic which is very inflectional? To deal with this question, we used root-stemming and light-stemming processes for all experiments conducted for stemming corpora and answers. We then measured the effect on the quality of the automatic grading.

### 4.1 COALS Semantic Space For Word Distribution

For semantic space creation we used the COALS(Correlated Occurrence Analogue to Lexical Semantic) algorithm (Rohde et al., 2004) for two main reasons.

*First*, it provides more consistent precision in the prediction of human similarity judgments than older algorithms such as Hyperspace Analogue to Language(HAL)(Lund and Burgess, 1996), Latent Semantic Analysis(LSA) (Deerwester et al. 1990 ), Random Indexing (Sahlgren, 2005). Second, unlike the algorithms that use sets of input documents (LSA for example which is widely used), COALS uses an undifferentiated text corpus and uses a moving window to define word collocations. The size of a COALS co-occurrence matrix is almost fixed, unlike an LSA matrix whose size is proportional to the number of documents. Therefore, COALS is proving to be much more *scalable and easy to implement*. This is suitable for Arabic language since finding a corpus can be a task with limited options due to the lack of Arabic resources(Al-Thubaity, 2015). For the purpose of these experiments it seemed most appropriate to use a diverse corpus of spoken Arabic language. This requires preprocessing on the corpus in order to prepare it for the extraction of the semantic space (Cleaning, Normalization, Stop-words Removal, Tokenization and Stemming). As shown in Fig. 3, the development pipeline of the semantic space highlights the generation of 3 matrices; co-occurrences matrix, correlation matrix and normalized correlation matrix into three steps:

a*) Gather co-occurrence counts using a ramped, size 4 window* compiling a co-occurrence matrix. Each element of the matrix represents the sum of the weights of the appearance of the row term with the column term, using the neighborhood according to the *distributional hypothesis*.

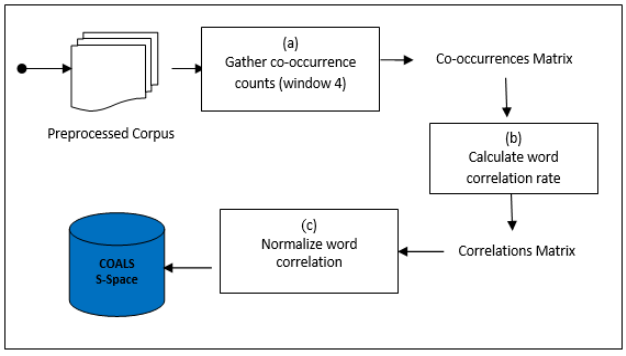


Figure 3: Semantic Space Creation Pipeline

[5] https://www.moodle.org/

[6] https://docs.moodle.org/37/en/Short-Answer_question_type

The hypothesis does not imply that similar words must appear next to one another, but that they should appear alongside the same set of other words. Concretely, if two words have neighborhood size smaller than the window, they are counted as coocurrents. With a window 4 we consider only the four adjacent neighbors on the left and on the right.

b) *Calculate the conditional co-occurrences rate.* The conditional co-occurrences rate aims to answer the question of whether a word (wi) occurs more or less often in the vicinity of another word (vj) than it does in general. To express this *tendency to co-occur*, Pearson's correlation is computed between the occurrences of words. Correlation matrix is then calculated by applying the formula (1) for each element of the co-occurrences matrix.

$$\begin{cases} w'_{a,b} = \dfrac{T w_{a,b} - \sum j\, w_{a,j} . \sum i\, w_{i,b}}{\left(\sum j\, w_{a,j} - (T - \sum j\, w_{a,j}) . \sum i\, w_{i,b} . (T - \sum i\, w_{i,b})\right)^{1/2}} \\ T = \sum_i \sum_j w_{i,j} \end{cases} \quad (1)$$

***a*** **&** ***b*** : Terms of the co-occurrence matrix (row i and column j).
$w_{a,b}$ : Element of the co-occurrence matrix of the terms a and b
$\sum j\, w_{a,j}$ : Sum of the columns of the term row a.
$\sum i\, w_{i,b}$ : Sum of the rows in the column of the term b
$T = \sum_i \sum_j w_{i,j}$ : Sum of all elements of the co-occurrence matrix.

Using this correlation, the new cell values will range from -1 to 1. A correlation of 0 means that term $w_i$ and term $v_j$ are uncorrelated and word $w_i$ is no more or less likely to occur in the neighborhood of $v_j$. A positive correlation means that term $w_i$ is more likely to occur in the presence of term $v_j$ than it would otherwise. For a large corpus, the correlation values are small, so it is rare for the correlation value to exceed 0.01. In addition, the majority of correlations are negative.

c) *Normalize the correlation matrix.* Negative values are normalized to 0(Negative correlations carry very little information) while positive values take their square root to amplify the importance of many small values relative to large values. Being symmetric, each row (or column) constitutes the semantic context vector of the row term (column term). All vectors of all words represent the semantic space. The generated semantic space is saved in a textual database. Context vectors are saved as long-string variables (LONGTEXT).

## 4.2 Arabic Stemming

The Arabic language belongs to the Semitic family of languages which also includes Hebrew and Aramaic. It is characterized by a lexicon built mainly from triliterous and quadrilateral roots, from a right-to-left writing system and from an alphabet composed of consonants. A stemmer is an automatic process in which morphological variants of terms are mapped to a single representative string called a stem(Lovins, 1963). The techniques used to proceed with stemming are generally based on a list of affixes (suffixes, prefixes) and on a set of rules of de-suffixation constructed a priori which allow, given a word, to find its stem. For Arabic language, finding the root, or stem, of a word is challenging to automate. The root of a word often has a very abstract meaning which is not at an appropriate level for NLP. In addition, words in Arabic can be 'borrowed' from other contexts, increasing ambiguity, presenting a challenge to the mechanical interpretation of Arabic. The two most useful approaches to Arabic stemming are based on root-extraction and on light-extraction stem. The Root stemming process consists of the removal of well-known prefixes and suffixes to extract the root of a word and to identify the pattern in correspondence with the remaining word. Light stemming is a less complex process and is stopped on removing prefixes and suffixes, without trying to identify the root word. We conducted our experimental synthesis using the two stemming techniques, light stemming and root stemming to derive the effect on the short answer grading. We used KHOJA' Stemmer [7](Khoja and Garside, 1999) for root stemming and Tashaphyne[8](Zerrouki, 2010) for light stemming.

## 4.3 Term Weighting

Term weighting allows to distinguish the discriminating words in the corpus from those which are less. In addition to the IDF (Inverse Document Frequency)(Salton and Buckley, 1988), we combine NTFlog (Normalized TFlog).

- TFlogs calculation for corpus terms applying (1):

$$TFlog(w) = -\log(Wc/N) \quad (1)$$

Wc: Number of times the term W appears in the corpus

N: Total number of words in the corpus.

- TFlogs maximum normalization applying (2) :

$$NTFlog(w) = TFlog(w)/Max(TFlog) \quad (2)$$

By applying (1), the very frequent words will have a low TFlog value and the rare words will have a high TFlog value. Since the least frequent terms are considered as the most discriminating, the maximum normalization of TFlogs (formula (2)), allows the very frequent words to have a standardized weight close to 0 and least frequent words will have a weighting close to 1.

## 4.4 Answer-to-Answer Similarity

In the same way that the corpus has been pre-processed, the Model Answer (MA) and the Student Answer (SA) are too (cleaning, normalization and stemming). The Bag Of Words (BOW) approach is used to represent the input answers. For answer to answer similarity, we combine weighted vector summation model and syntactic similarity based on common words between MA and SA.

### 4.4.1 Vector Summation Model

Consists of summing the context vectors of each word for each answer and then calculating the cosine similarity between the sums of the vectors of the two answers (MA, SA). Let MA be composed of the words M1, M2 … MN Let SA be composed of the words K1, K2 … KM First, Extract from the semantic space all words vectors V(Mi), V(Ki)v($M_i$) and calculate the sum vectors answers VMA and VSA:

$$VMA = \sum V(Mi) * \beta i \quad (i=1, N)$$

[7] http://zeus.cs.pacificu.edu/shereen/research.htm#stemming

[8] https://pypi.org/project/Tashaphyne/

$$VSA = \sum V(Ki) * \beta i \quad (i=1, M)$$

$$Sim(MA, SA) = Cosine(VMA, VSA)$$

We used two varieties of this model by assigning two values to βi:

Un-Weighted Vector Summation Model (βi = 1) referred to Basic System (baseline).
Weighted Vector Summation model (βi = IDFi* NTFlogi) referred to W-SM Model in the following.

#### 4.4.2 Combined Similarity Model

Naturally in the student's answer there are probably several words in common with the model's answer. So the Dice coefficient(DICE, 2012) is combined to further favor cases with a significant number of common words between the two answers. It measures the syntactic similarity between two answers based on the number of their common terms.

Sim (MA, SA)=(2*Nc) /(NMA+NSA)

Nc: number of common terms to MA and SA,

NMA: number of MA terms and

NSA: number of MA terms.

To enhance more scores, an unsupervised combination which rely on a max of the scores between weighted Summation Model (IDF*NTFlog) and DICE's Coefficient is applied. This model referred to Proposed Model.

### 4.5 Scaling Grades

The Similarity takes value in [0..1]. In a grading task, the output must be an understandable grade that occurs in a well-defined interval of grades. The task of the scaling is then to map the similarity value to a grade. Unsupervised K-Means clustering ((MacQueen, 1967) is used. The idea behind using this clustering method is that all scores similarity are separated into k clusters, and each cluster is represented by its centroid (the sum of the grades divided by their number). After defining k centroids, each grade is assigned to a cluster by using the Euclidean distance d. Then, the centroids are recalculated until we find an optimal set of clusters. K is fixed to 11. Each cluster refers to one grade from all possible 11 grades obtained in annotation (0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5) in the scale [0..5].

## 5. Evaluation Results and Discussion

In this section, the AR-ASAG Dataset is used to evaluate the performance of the proposed grading approach. An analysis and the comparison of the results obtained are presented and discussed with four perspectives: Semantic space dimensionality and Domain specificity, semantic space quality vs. Word Embedding distribution, Grading Model Quality Assessment with term weighting and stemming impact.

**Evaluation metrics.** The analysis of several evaluation dimensions, imposes a coherent choice of metrics. This was influenced by related work that used the same Datasets. *Pearson correlation* (r: the higher the better) is the most frequently used metric for research in this area. We reported it for all our experiments. Although it is not cited and used in the majority of related work, we use jointly with the Pearson coefficient, the *Root Mean Squared Error* (RMSE, the lower the better). In the following, we present the results of experiments for 10 answers for each question. To make a correct sampling, a randomized split, mixes the Dataset before the selection. Then for each question, the first ten answers have been selected to obtain a sub dataset of 480 p (48*10) of (model answer, student answer) including all question types presented in section 3.1. For reproducibility, this Dataset subset is named in the complete dataset as AR-ASAG-480. Both answers and corpora are preprocessed and stemmed.

### 5.1 Semantic space dimension and domain specificity

When applying corpus-based techniques, one of the key things to consider is the extent to which size and subject affect the overall performance of the system. In particular, based on the underlying processes involved, the COALS algorithm should be particularly sensitive to changes in domain and semantic space dimension. Semantic measurement depends on the correlation of words in the learning corpus, suggesting that, for example, in the field of cyber-crimes, the terms "crime" and "internet" will be more closely related than in a more general text corpus. Naturally, a large amount of learning data will result in a reduction of vector spaces, which in turn should affect the performance of corpus-based models.

***In-domain CYBER Corpus***. Because of unavailability of such specific corpus in Arabic, we developed our own Corpus *(Arabic Cyber Text Corpus, 2020)* covering the field of cybercrimes. The domain-specific corpus was automatically obtained from texts extracted from a collection of URLs according to a list of key terms. Key terms are combined and queried to a search engine, which returns a list of potentially relevant URLs. The URLs are then inspected and validated. Relevant web pages are retrieved, automatically cleaned of HTML tags. The text is extracted and added to the corpus. The corpus was then enriched by several course notes covering the topics used as questions in AR-ASAG Dataset. We also use three publicly available generic corpora (BBC Arabic, CNN Arabic)[9], and Khaleej[10]. Table 4 resumes characteristics of corpora used to build different semantic spaces.

| | BBC | CNN Arabic | Khaleej | CYBER |
|---|---|---|---|---|
| Words | 1 860 000 | 2 241 348 | 3 000 000 | 2009110 |
| Documents | 4763 | 5690 | 5000 | 1273 |
| Contents | - Middle East News<br>- Economy &work<br>- Sports<br>- International press<br>- Science & technology<br>- Arts and cultures | - News of the world<br>- Economy and works<br>- Sports<br>- Science & technology<br>- Arts and cultures | - International News<br>- Local News<br>- Sports<br>- Economy | - Cyber crimes<br>- Information Security culture<br>- Cyber crimes classification<br>- Cyber crime legislation |

Table 4: Corpora Characteristics

The Singular Value Decomposition (SVD) algebra method is an important tool for factorizing complex rectangular matrices to reduce size (the semantic space is a matrix). However, this method is very expensive in terms of memory consumption and may be unworkable for larger corpora in

[9]https://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/

[10] https://sites.google.com/site/mouradabbas9/corpora

particular, where initial space dimension may be important. Ideally for our proposed system, the SVD algorithm would be computed using the full matrix word vectors of the built semantic space. However, this is computationally difficult and is unnecessary. Good results can be obtained using several thousand of the most frequent words. The semantic space dimensionality can be reduced by increasing the limit of the infrequent words. In Table 5, we explore four semantic spaces with dimensionality ranging from 13733 to 28062. In the following, evaluation is done with the basic system on the AR-ASAG Dataset. Basic system refers to the system using the Un-Weighted summation model. The baseline assigns a grade based on the cosine similarity between the vector space of the model answer and the student answer.

| | *Root Stemming* | | | *Light Stemming* | | |
|---|---|---|---|---|---|---|
| **Semantic Space** | **Vector Dimension** | **Pearson** | **RMSE** | **Vector Dimension** | **Pearson** | **RMSE** |
| **khaleej** | 13733 | 0.6306 | 1.13 | 18630 | 0.6087 | 1.21 |
| **cnn** | 16752 | 0.6317 | 1.14 | 21032 | 0.6090 | 1.20 |
| **bbc+cnn** | 24230 | 0.6379 | 1.14 | 28062 | 0.6115 | 1.22 |
| **CYBER** | 17225 | **0.6550** | **1.10** | 23715 | **0.6340** | **1.14** |

Table 5: Basic system results for different semantic spaces on AR-ASAG Dataset

For each corpus, several semantic spaces were generated by varying the limits of the most frequent words. We present in Table 5, for each corpus, the semantic space that gave the best results. By increasing dimensionality, the basic system presents the best result with dimension 17225 using a light stemming and 23715 using a root stemming. Above, there is a little change in performance. Performance declines slowly as we reduce the vectors dimension to 13000. This well confirms that in practice roughly equivalent performance is obtained from using anywhere dimensionality from 14,000 to 100,000 using COALS algorithm(Rohde et al., 2004). Considering the Khaleej space which has comparable dimension with CYBER, we notice that the CYBER Space is more efficient for both stemming techniques. Comparing obtained results, we see that by using the in-domain CYBER space we obtain a correlation of r=0.6550 and an RMSE =1.10, which is higher than the correlation of r=0.6379 and RMSE = 1.14 obtained with a corpus of bigger dimension oriented to general domain. This suggests that for COALS algorithm, the quality of the texts is more important than their quantity. This result has a double advantage for the proposed approach. First, encourages the use of specific domain corpora since around medium dimensionality results are better. Second, it is more easy to build (or find) any corpus (not necessarily of gigantic size) and to not require a lot of machine resources for the implementation of the grading system. All the results reported in the following sections are computed using the CYBER semantic space.

## 5.2 Word Space Distribution Quality vs. WE

Authors in (Zahran et al., 2015) used CBOW, SKIP-G and GloVe models(Mikolov et al., 2013) to build a multidimensional word representation in vector space for Modern Standard Arabic. This model of Word Embedding(WE) is about 6.3 million entries and the total number of words is about 5.8 billion. Vectors of (Zahran et al., 2015) are referred here by Zahran-WE.

Taking advantage of the availability of Zahran-WE(Zahran et al., 2015), we evaluated word distribution in the semantic space vs. WE distribution. Thus, in our basic system, we replaced the semantic space vector words by Zahran-WE (from the CBOW and SkipGram models) and calculated the correlation on the AR-ASAG Dataset. These systems are referred to Z-CBOW Basic and Z-SkipGram Basic in the following. In table 6, we can find the results of the confrontation with basic systems.

| | *Root Stemming* | | *Light Stemming* | |
|---|---|---|---|---|
| **Basic System using** | **Pearson** | **RMSE** | **Pearson** | **RMSE** |
| Z-CBOW WE | 0.6433 | 1,13 | **0.6475** | 1,14 |
| Z-SkipGram WE | 0.6281 | 1,30 | 0.6348 | 1,22 |
| CYBER Space | **0.6550** | **1.10** | 0.6340 | **1.13** |

Table 6: Basic system performance using WE vs. Cyber semantic space on AR-ASAG Dataset

Basic system using the CYBER semantic space, records a similar performance with Z-CBOW basic (Pearson -0.01, RMSE + 0.01) while it outperforms Z-SkipGram Basic (*root stemming*: Pearson +0.03, RMSE +0.2; *light stemming*: Pearson -0.00008, RMSE +0.09). This finding gives a good indication of the quality of the word distribution in semantic space. Note that for the Zahran-WE generation, Authors have trained non-stemmed text corpora. what could explain the slight difference for obtained results by Z-CBOW with the light stemming (of answers).

## 5.3 Grading System Assessment

We discuss here the quality of the proposed grading model on 2 dimensions. First, with respect to the human correlation on the AR-ASAG Dataset discussing term weighting and stemming effect. Second, in relation to the results of the competition SEMEVAL 2017 (track 1: Arabic-Arabic) on the STS 250 Dataset.

### 5.3.1 Term Weighting Effect

in Table 7, we can find performance of proposed approach against Inter Annotator Agreement on AR-ASAG Dataset. The term weighting gave an interesting correlation improvement *(root stemming: Pearson +0.028, RMSE +0.04; light stemming: Pearson +0.0478, RMSE +0.07)*. The term weighting combined with the syntactic measure put the system in its best correlation **(r=0.7037).** The RMSE was markedly improved to **1.0240** (RMSE +0.14).

| | | **Pearson** | **RMSE** | **Av(RMSE)** |
|---|---|---|---|---|
| **IAA ( Manual scores )** | | 0.8384 | 0.8381 | *0,5629* |
| **Basic System (Arabic)** | *Root Stem.* | 0.6550 | 1.10 | |
| | *Light Stem.* | 0.6340 | 1.14 | |
| **W-SM System (Arabic)** | *Root Stem.* | **0.6830** | **1.06** | |
| | *Light Stem.* | 0.6818 | 1.07 | |
| **Combined Proposed System** | *Root Stem.* | 0.7010 | **1.0240** | ***0.7841*** |
| | *Light Stem.* | **0.7037** | 1.0454 | ***0,8039*** |

Table 7: Proposed system evaluation final results on AR-ASAG Dataset

### 5.3.2 Stemming Effect

As we can see in Tables 5, 6 and 7, root stemming gave better results than light stemming in basic system. But surprisingly when combining term weighting and syntactic

similarity, the light stemming gave comparable results (+0.0027) with a lesser RMSE (-0.0214). In addition, we report the RMSE error for the Dataset and the median error RMSE for each question. This gives an indication of the performance of the system allowing a single question to be noted in isolation. The average RMSE error for the root stemming (0.7841) is then better than The average RMSE error for the light stemming (0.8039) but with a small difference. The average RMSE confirms a slight difference for root stemming although Pearson announces the opposite. This is visible on the distribution of manual and automatic grades for both stemming techniques in Fig. 4, where the two relating curves have the same tendency and are almost confused. The basic system correlation was very sensitive to the stemming technique. The results remained comparable with term weighting and combined similarity. Our experiments have shown that the Root and Light stemmers perform automatic grading tasks with statistically equivalent correlation with human grades. We found that overall a light stemmer performs automatic grading as well as complex root stemmers. The effect of the combined similarity model is more relevant than the stemming technique itself. This implies that with less ***complex*** light stemming techniques, the results remain comparable when especially root stemming tools are far from having reached maturity in Arabic NLP tasks. This

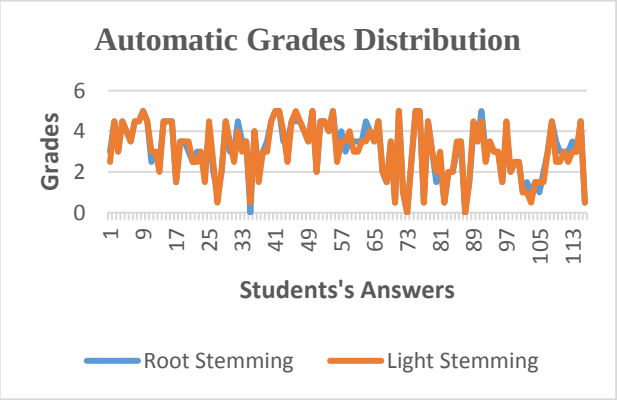


Figure 4: Automatic Grades Distribution Root Stemming vs. Light Stemming

is very interesting for the Arabic language, which although great efforts led by different researchers, the lack of tools and resources still remains a real challenge.

***Overall Grading assessment.*** A thorough analysis of automatic grades compared to human correlation allows us, as shown in Table 8, to better appreciate the system's performance despite a Pearson Correlation. We choose final results using root stemming for the in-domain space for the analysis.

| Difference | Manual - Manual | | Manual - Automatic | |
|---|---|---|---|---|
| | *Number of answers* | *%* | *Number of answers* | *%* |
| 0 | 743 | 34.83 | 252 | 11.81 |
| 0 <D <= 1 | 1155 | **54.14** | 1252 | **58.69** |
| 1 <D <= 2 | 189 | 8.86 | 475 | 22.26 |
| 2 <D <= 3 | 38 | 1.78 | 130 | 6.09 |
| D > 3 | 8 | 0.37 | 24 | 1.12 |

Table 8: Manual-Manual and Manual-Automatic Grades Analysis on AR-ASAG dataset

The difference between manual-manual and manual-automatic grades are comparable. Effectively, in 58.69%, the manual-automatic difference is between 0 and 1 which is better compared to 54.14% of manual-manual difference. In 70.5% (58.69 + 11.81) the manual-automatic difference is less than or equal to 1. This gives a good indication for the proposed model. 92.76% of answers have a variance of up to 2 on a scale of 5 points. This is reasonable enough by means of the subjectivity of the evaluation process itself. In 7.21%, the manual-automatic difference is strictly greater than 2. Compared to a manual-manual difference of only 2.15%, this manual-automatic difference must be reduced in future versions of the system.

***Comparing on STS 250 SEMEval Dataset.*** In Table 9, we report the results obtained on generic STS 250 SEMEval 2017 dataset. Proposed system achieved 11,75 % higher than the SEMEVAL baseline but achieved -2,43 % lower than LIM-LIG (Nagoudi and Ferrero, 2017) the second score in track 1, that used a vectorized Word Embedding based approach(similar to ours) and 3.23 % lower than the track1'winner(Huang and Su, 2017) using a topological approach. As the RMSE is not mentioned in the SEMEval competition we compared the obtained grades to the manual ones. Indeed, over 250 answers, 69,6 % present a difference less than or equal to 1 and 92,4 % present a difference less than or equal to 2. In 7.6%, the manual-automatic difference is strictly greater than 2. From the results on STS 250 SEMEval dataset, that we consider acceptable, we can learn two things. First, the proposed approach can well generalize. Second, an interesting indication on the quality of the AR-ASAG Dataset since the same system operates in a comparable way on two different Datasets: (r: 0.7037, RMSE: 1.0240) on AR-ASAG and (r: 0.7220, RMSE: 1.03) on SEMEval Dataset with a same RMSE.

| | Pearson | RMSE |
|---|---|---|
| **SEMEVAL 2017 track 1 Baseline** | *0.6045* | - |
| **SEMEval 2017 Winner Track 1 BIT System** (Nagoudi and Ferrero, 2017) | ***0.7543*** | - |
| **SEMEval 2017 2nd score Track 1 LIM-LIG** (Huang and Su, 2017) | **0.7463** | - |
| **Our Basic System** | *0.6303* | 1,31 |
| **W-SM Model** | *0.6801* | 1,19 |
| **Proposed System** | ***0.7220*** | **1.03** |

Table 9: Proposed System Evaluation Final Results on STS 250 SEMEval 2017 Dataset (Task 1)

## 6. Conclusions

In this paper, we introduced AR-ASAG, an Arabic Dataset for Automatic Short Answer Grading and then we explored an unsupervised vector space approach for automatic grading which is evaluated using the Dataset to serve as baselines for future research work. We believe the paper made three important contributions.

*First*, in order to stimulate research in Arabic language, there is a dire need to develop Datasets in this language. To our knowledge, AR-ASAG is the first Arabic Dataset for automatic short answer grading publicly available for download in Arabic Language. We believe that it will be a valuable evaluation contribution as more researchers will

reuse the publicly available Dataset as opposed to using data from restricted internal sources.

*Second*, the proposed grading approach is particularly suitable for languages that face the challenge of lack of tools and resources since the only language-dependent resources are a corpus and a stemmer.

*Third*, there are different things that we can learn from the experiments conducted. Indeed, improvements can be obtained when using a medium size domain-specific COALS semantic space unlike a semantic latent analysis that advocates large corpora. This suggests that for COALS algorithm, text quality is more important than their quantity. The term weighting combined with the syntactic measure put the system in its best correlation maintaining the system simple and feasible in practice. The effect of term weighting and syntactic combined similarity is more relevant than the stemming technique as in overall a light stemmer performs as well as complex root stemmers. This is especially interesting for Arabic Language where root stemmers are far from having reached maturity. Evaluated on Dataset, the approach used in this work gives promising results for Arabic language. It gets significantly closer correlation to human grades on AR-ASAG Dataset. It approaches some results in the literature. The proposed approach would apply not only to Arabic language but also to others having similar challenges. The proposed system is implemented as integrated plug-in on the Moodle LMS platform. Work is underway for a quantitate and a qualitative scalability evaluation. In future work, we concentrate on improving the quality of the answer grading by training a supervised model to consider more input from the teachers, and more features that correctly mirror real-world issues associated with the task of grading. As the approach used is language independent it can be tested and implemented for other languages in near future.

## 7. Acknowledgements

This work is supported by the Ministry of Higher Education and Scientific Research in Algeria (Project C00L07UN100120180002). Authors are grateful to Faiza OUKINA and Imene AMARSETTI for their technical help.

## 8. Bibliographical References

Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, *49*(3), 721–751. https://doi.org/10.1007/s10579-014-9284-1

Anderson, Lorin W.; Krathwohl, David R.; Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives* (omplete ed., Ed.). New York : Longman.

Bennouar, D. (2017). An Automatic Grading System Based on Dynamic Corpora. *The International Arab Journal of Information Technology*, *14*(4A).

Brown, Sally, Ed.; Glasner, Angela, E. (1999). Assessment Matters in Higher Education: Choosing and Using Diverse Approaches., 1999.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. In *International Journal of Artificial Intelligence in Education* (Vol. 25). https://doi.org/10.1007/s40593-014-0026-8

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. In Association for Computational Linguistics (Ed.), *Proceedings ofthe 11th International Workshop on Semantic Evaluations (SemEval-2017)* (pp. 1–14). https://doi.org/10.18653/v1/S17-2001

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (n.d.). *Indexing by Latent Semantic Analysis*.

El Moatez Billah Nagoudi, Jérémy Ferrero, D. S. (2017). LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting. In Association for Computational Linguistics (Ed.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 134–138).

Gomaa, W. H., & Fahmy, A. A. (2014a). Arabic Short Answer Scoring with Effective Feedback for Students. *International Journal of Computer Applications*, *86*(2), 35–41. https://doi.org/10.5120/14961-3177

Gomaa, W. H., & Fahmy, A. A. (2014b). Automatic scoring for answers to Arabic test questions. *Computer Speech and Language*, *28*(4), 833–857. https://doi.org/10.1016/j.csl.2013.10.005

Gratta, R. del, Frontini, F., Fahad, K., Mariani, J., & Soria, C. (2014). *The LREMap for under-resourced languages*.

J. B. MacQueen. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Khoja, S., & Garside, R. (1999). Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

LEE R. DICE. (2012). *Measures of the Amount of Ecologic Association Between Species Author ( s ): Lee R . Dice Reviewed work ( s ): Published by : Ecological Society of America Stable URL : http://www.jstor.org/stable/1932409 . 26*(3), 297–302.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12028

Lovins, J. (1963). *Development of a stemming algorithm.* Cambridge: M.I.T. Information Processing Group Electronic Systems Laboratory.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods,*

*Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Mag`ooda, A., Zahran, M. A., Rashwan, M., Raafat, H., & Fayek, M. B. (2016). Vector Based Techniques for Short Answer Grading. *FLAIRS Conference*, 238–243.

Mahmoud El-Haj, Udo Kruschwitz, Colchester, & Fox, C. (2015). Creating Language Resources for Under-resourced Languages : Methodologies , and experiments with Arabic. *Language Resources and Evaluation*, *49*(3), 549–580. https://doi.org/10.1007/s10579-014-9274-3

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 775–780. https://doi.org/10.1.1.65.3690

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Retrieved from http://arxiv.org/abs/1301.3781

Mohler, M., Bunescu, R., & Mihalcea, R. (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*, 752–762.

Mohler, M., & Mihalcea, R. (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 on - EACL '09*, 567–575. https://doi.org/10.3115/1609067.1609130

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2004). An Improved Method for Deriving Word Meaning from Lexical. *Cognitive Psychology*, *7*, 573–605.

Sahlgren, M., (2005). An introduction to random indexing. In Methods and Applications of Semantic Indexing Worshop at the 7th International Conference on Terminology and Knowldge Engineering, TKE 2005.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Shourya Roy, Y. Narahari, and O. D. D. (2015). A Perspective on Computer Assisted Assessment Techniques for Short Free-text Answers. *Communications in Computer and Information Science*, *571*(June 2015). https://doi.org/10.1007/978-3-319-27704-2

Wu, H., Huang, H., Jian, P., Guo, Y., & Su, C. (2017). BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity. In Association for Computational Linguistics (Ed.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*: *Vol. L* (pp. 77–84).

Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H., Rashwan, M., & Atyia, A. (2015). Word Representations in Vector Space and their Applications for Arabic. In A. Gelbukh (Ed.) (Ed.), *16th international conference, CICLing 2015 Cairo, Egypt, april 14* (Vol. 9041, pp. 430–443). https://doi.org/10.1007/978-3-319-18111-0_32

## 9. Language Resource References

(AR-ASAG Dataset, 2020). The Arabic Dataset for Automatic Short Answer Grading Evaluation, V. 1.0, ISLRN 529-005-230-448-6.

(Arabic Cyber Text Corpus, 2020). The Arabic In-Domain Cyber Text Corpus, V. 1.0, ISLRN 798-080-268-332-8.