Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Evaluating Pretrained Embeddings for Automatic Short Answer Grading

Nikunj C. Gamit [1*], Dr. Shailesh Panchal [2]

[1] Research Scholar, Gujarat Technological University, Gujarat, India - 382424
[2] Professor, Graduate School of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India - 382424
**Corresponding Author: nikunjgamit@gmail.com*

ARTICLE INFO

Received: 19 Sep 2023
Accepted: 26 Nov 2023

ABSTRACT

Automatic Short Answer Grading (ASAG) refers to the computer methods employed to assess student responses based on a given question and the expected answer. Prior studies employed various techniques such as idea mapping, facet mapping, and standard word embeddings to extract semantic information. The researchers manually retrieved different features in order to train the models using the respective datasets. In this study, we evaluate the effectiveness of pretrained embeddings using transfer learning models, including Word2Vec, USE, BERT, and GloVe for the purpose of assessing their performance for ASAG. Our training process involves utilizing cosine similarity derived from the embeddings of the aforementioned models. The error scores and correlation values of the four models are compared with those of prior studies on the Mohler dataset. The findings of our study indicate that the USE model exhibited superior performance compared to the other three models. An examination of potential factors contributing to suboptimal performance of such models is analyzed.

Keywords: Short answer grading, semantic similarity, automatic assessment, Word2Vec, USE, BERT, GloVe

INTRODUCTION

Descriptive assessments measure the extent of pupils' comprehension of various topics or subjects. The increasing enrolment of students in online platforms and colleges has become a notable trend. Assessing all the responses might be a demanding task when done through manual effort and without the aid of automation or machinery. The procedure for evaluating these comprehensive responses is accomplished through the utilization of Automatic Short Answer Grading (ASAG) methods. The ASAG approaches assesses the performance of students' answers by comparing them with given reference answer(s).

[1] and [2], used corpus-based approaches for the assessment of student answers. [3], [4], and [5] integrated machine learning techniques for the evaluation purpose. The techniques for obtaining syntactic and lexical information as well as identification and analysis of morphological and semantic elements into ASAG has been explored by [6]. In prior studies, researchers have manually integrated certain features. Semantic similarity is one of such features. These linguistic characteristics are then utilized for the purpose of training a classifier or regression model. These are considered as a task of finding Semantic Textual Similarity (STS) as well. STS involves assigning a measure of similarity between texts. The objective is to compare the scores of two textual corpora. The most recent developments in the field of Natural Language Processing (NLP) and deep learning have proven to be promising in their contributions in the field of ASAG. The work by [7] and [8] introduced novel architectures in the field of transfer learning. These models were trained on huge corpora and possess the capability to derive the semantic context of the words. As we discuss further, these models include Word2Vec [9], GloVe [10], Bidirectional Encoder Representations from Transformers (BERT) [11], and Universal Sentence Encoder (USE) [12].

The embeddings of the transfer learning models are utilized to extract semantic knowledge from the answers through encoding with contextual vectors. In contrast to prior studies, our approach involves a reduced number of pre-processing steps while training a regression model, in order to evaluate the efficacy of the models. The efficacy of the embeddings of these models is assessed using the Mohler dataset [6]. This study examines the efficacy of utilizing embeddings derived from pretrained transfer learning models for the purpose of ASAG.

RELATED WORK

The researchers [13], [14], and [15] worked with concept mapping methodologies, which involved mapping the relevant concepts of student answer with intended answer. Pattern matching was used in further research by [16], [17], and [18] to extract information from student answers. They were able to extract the patterns by utilizing regular expressions in conjunction with parse trees. [1] gave ASAG its first exposure to corpus-based approaches by combining scores from Latent Semantic Analysis (LSA) and Bi-lingual Evaluation Understudy (BLEU). [19] utilized the combination of several knowledge-based and corpus-based features for the purpose of extracting the similarity between the students' answer and the teacher's answer. The next development incorporated corpus-based approaches into machine learning systems [6].

In their study, [20] employed various features like word-alignment, vector similarity, and term frequency–inverse document frequency (tf-idf). These features were trained and evaluated using the Mohler dataset, as previously introduced by [6]. [21] employed established natural language processing (NLP) techniques, including Word2Vec [9], [22], GloVe [10], and FastText [23], to extract distributional and semantic information via embeddings. They conducted a comparative analysis of the outcomes achieved by employing pretrained word embeddings and domain-specific trained word embeddings from Word2Vec, GloVe, and FastText models. The evaluation was carried out using the Mohler dataset. The traditional embeddings failed to take into account the contextual nuances of words and the presence of long-term dependencies. Subsequent developments in Natural Language Processing (NLP) have involved the consideration of contextual information pertaining to words within phrases, which has been facilitated by pretraining models on extensive corpora. As a consequence, various models were developed, such as InferSent [24], ELMo [25], GPT [26], BERT [11], and GPT-2 [27]. Moreover, the previous studies commonly employ a variety of criteria to assess the responses. We exclusively utilized semantic similarity attribute, derived from the transfer learning models, for the purpose of training.

A summary of transfer learning models considered for experimentation purpose is given in Table 1.

Table 1. An overview of models used for experimentation

Model	Architecture	Dataset	Dataset Size
USE	Transformer	Multi-domain dataset	Not known
Word2Vec	CBOW / Skip-gram	Google News	100B Words
BERT	Stacked transformer	BookCorpus, Wikipedia	800M words, 2500M words
GloVe	NA	Crawl	840B tokens

DATASET

The Mohler dataset consists of a collection of questions and corresponding responses within the field of computer science, as described by [6]. The objective of the dataset is to assess the model's performance in assessing students' responses by comparing them to the desired answer provided by the evaluator. The dataset comprises a total of 2273 responses, which were obtained from 31 students who completed 10 assignments and 2 tests. These responses were collected for a set of 80 distinct questions. The assignment's answers are assessed on a scale of 0 (indicating incorrect) to 5 (indicating complete accuracy) by two evaluators who possess expertise in the domain of computer science. The standard score of each answer is determined by calculating the average of the scores given by the two assessors. The responses provided in examinations were evaluated on a scale ranging from 0, indicating an incorrect answer, to 10, indicating a completely accurate response. In order to reduce the presence of assigned scores inside the dataset, [21] implemented a normalization process to standardize all examination grades on a scale of 0 to 5. According to [6], the dataset exhibits a bias towards accurate responses. The average grade is reported to be 4.17, whereas the median grade is 4.50.

EXPERIMENTS

Experimental Setup: The experiments were run on Google Colab Pro, using its cloud-based GPU resources for the support of training and testing different embedding models that are applicable to Automatic Short Answer Grading (ASAG). The computation environment comprised a Tesla T4 GPU (or equivalent graphics processing unit, as and when available), 16 GB RAM, and CUDA Version 11.x, all of which operated in the framework of an Ubuntu-based Google Colab environment. The configuration was able to provide adequate processing power to accommodate both traditional and deep learning-based models in an efficient manner. To test and compare various methods of text representation, a diverse set of Python libraries and frameworks were utilized. Gensim facilitated training and utilization of Word2Vec, FastText, and GloVe embeddings, and Scikit-learn enabled running Bag of Words (BoW) and TF-IDF models, together with calculation of a variety of evaluation metrics. Pre-trained transformer-based embeddings, such as BERT and Universal Sentence Encoder (USE), were utilized using the Hugging Face Transformers library for the efficient utilization of contextual word and sentence embeddings. The Sentence-Transformers library was utilized for fine-tuning embeddings at the sentence level for applicability in short answer marking tasks. The Torch (PyTorch) was also utilized as the deep learning framework required to train transformer-based models, whereas the Hugging Face Datasets library was utilized to efficiently preprocess the data. These tools, integrated together, made a comprehensive test of a range of embedding strategies feasible, trying their efficacy on the ASAG domain.

Preprocessing & Feature Extraction: In the preprocessing stage, we have solely performed tokenization. The deliberate omission of lemmatization and stop word removal is undertaken in order to evaluate the efficacy of transfer learning models in their unprocessed state. [21] employed a spell checker to rectify misspelled words, a practice that we have refrained from utilizing. It has been posited that the graders have imposed deductions on the pupils' answers for misspelled words. In this regard, the presence of spelling errors might be regarded as a detrimental attribute that can be internally addressed by training. Given that transfer learning models are trained on extensive vocabularies, it is reasonable to hypothesize that they possess a certain level of capability in comprehending misspelled words. The adaptability of transfer learning models in generating embeddings for novel words is considered to have aided in discarding spelling errors.

The tokens of each word in all the answers are assigned the pretrained embeddings of each transfer learning model. Given the existence of several words that can serve as an answer to a particular question, we generate answer embeddings using the average of word embeddings, as illustrated in Equation 1. In this equation, a_{ij} denotes the vector representation of the j^{th} answer for question q_i, while w_k and n_j represents the vector representation of the k^{th} word and number of words in the answer a_{ij}, respectively. This process generates a singular vector that represents each response within a high-dimensional hypothesis space. The dimensions of the sentence embeddings are equivalent to the dimensions of the word embeddings.

$$a_{ij} = (\sum_{k=1}^{n_j} w_k)/\, n_j \tag{1}$$

The cosine similarity, as defined in Equation 2, is utilized to compute the similarity between each student answer a_{ij} and the intended answer a_i. The scores are normalized on a scale of 0 to 1 in order to standardize the similarities and obtain a relative measure of similarity. The scores are regarded as the features of the answers and are trained using various regression algorithms.

$$\cos(a_{ij}, a_i) = \frac{a_{ij}.a_i}{|a_{ij}||a_i|} \tag{2}$$

Training and Testing: The Mohler data was divided into two subsets, with 65% of the data used for training purposes and the remaining 35% reserved for testing. Each model is trained for 1000 iterations, with the training and testing data being randomly selected for each iteration to ensure the generalizability of the results. The cosine similarity feature is trained using linear, polynomial, and Lasso regression, with the corresponding grades being assigned. The regression models chosen for implementation are utilized to conduct a comparative analysis of our findings with those presented in the studies conducted by [6] and [21]. After completing the training process, the regression model is evaluated using test data. The test data remains unseen by the regression model until it enters the testing phase. The similarity scores of the test data are fed into the trained regression model as input. As a consequence, the anticipated grades will be subsequently utilized for assessment. During the process of evaluation, it is customary to compute the Root Mean Square Error (RMSE) and Pearson correlation coefficient to assess the relationship between the anticipated scores and the desired values.

RESULTS

Table 2 presents the RMSE and Pearson correlation coefficient (ρ) values obtained from the utilization of pretrained embeddings from transfer learning models on the Mohler dataset. The RMSE score quantifies the magnitude of the discrepancy between the anticipated and projected results. Consequently, a model's performance improves as the RMSE decreases. The Pearson correlation coefficient quantifies the extent of the overall agreement between the allocation of intended scores and projected scores. Consequently, a larger value of ρ indicates a superior performance of the model. Table 3 presents a comparative analysis of our findings in relation to previous methodologies and models.

Table 2: RMSE and Pearson correlation (ρ) of pretrained TL models on Mohler Dataset

Model	Polynomial regression		Linear regression		Lasso regression	
	RMSE	ρ	RMSE	ρ	RMSE	ρ
USE	0.997	0.561	0.988	0.461	0.989	0.452
GloVe	1.081	0.239	1.074	0.241	1.083	0.233
BERT	1.062	0.309	1.066	0.272	1.085	0.271
Word2Vec	1.289	0.218	1.219	0.209	1.159	0.193

For USE, the embeddings obtained Pearson correlation coefficient of 0.561 and RMSE score of 0.997 when used in conjunction with Polynomial regression. The Pearson correlation coefficients for Linear regression and Lasso regression are 0.461 and 0.452, respectively. The RMSE scores for Linear regression and Lasso regression are 0.988 and 0.989, respectively. The utilization of GloVe embeddings yielded Pearson correlation coefficients of 0.239, 0.241, and 0.233. The RMSE values obtained from the analysis of GloVe embeddings are 1.081, 1.074, and 1.083.

Table 3: Overview comparison of results on Mohler dataset with former approaches

Model	Features	RMSE	ρ
BOW	SVM Rank	1.042	0.480
	SVR	0.999	0.431
tf-idf	SVR	1.122	0.327
tf-idf	LR + SIM	0.923	0.591
FastText	SOWE + Verb phrases	1.023	0.465
	SIM + Verb phrases	0.956	0.501
Word2Vec	SOWE + Verb phrases	1.289	0.218
GloVe	SOWE + Verb phrases	1.081	0.239
USE	SIM	0.997	0.561
BERT	SIM	1.062	0.309

The BERT embeddings exhibited better performance when employed in conjunction with Polynomial regression, yielding a correlation coefficient (ρ) of 0.309 and RMSE of 1.062. In comparison, Linear regression produced a correlation coefficient of 0.272 and an RMSE of 1.066, while Lasso regression resulted in a correlation coefficient of 0.271 and an RMSE of 1.085. In terms of performance, Word2Vec has demonstrated poor results compared to other models across three regression training models. The respective ρ values for Word2Vec were 0.218, 0.209, and 0.193, while the corresponding RMSE scores were 1.289, 1.219, and 1.159.

Table 3 presents a comparison between our findings and the traditional embeddings of BOW, tf-idf and FastText using the Mohler dataset. However, we exclusively focus on models or approaches that refrain from utilizing domain-specific training of the data in order to ensure a fair and comprehensible comparison. The Features column of Table 3 lists the diverse range of features or methods employed by the models. The findings presented by [20] demonstrate that the tf-idf technique, incorporating both length ratio and similarity features, had best performance score as compared to alternative approaches, as indicated by a Pearson correlation coefficient of 0.591 and a RMSE of 0.923.

DISCUSSION

The findings of our study demonstrate that the USE model had superior performance in the context of domain-specific ASAG when compared to other transfer learning models. The primary factor that may account for the superior performance of USE compared to other transfer learning models on the dataset particular to the given domain is as follows. The fact that USE requires same embeddings to work on multiple generic tasks to capture most generic information. These generic embeddings are suitable to multiple tasks such as relatedness, clustering, text classification and paraphrase detection. The pretrained datasets of BERT, Word2Vec, and GloVe models are substantial, but the domain we are testing is relatively smaller in scale. As a consequence, the similarity scores ranged from 10^{-5} to 10^{-1}.

The performance of Polynomial regression was better than that of Linear and Lasso regressions. The reason for this is that Polynomial regression employs a step-wise training approach, which has resemblance to the manual assignment of grades to students, in contrast to linear and non-linear regression techniques. When compared to previous methodologies, the utilization of Universal Sentence Encoder (USE) embeddings stands out due to the implementation of several preprocessing techniques, including multiple steps for feature extraction and training.

CONCLUSION

We performed an evaluation of the embeddings generated by four transfer learning models on the Mohler dataset, which is peculiar to a certain domain. The evaluation was performed on the task of Automatic Short Answer Grading (ASAG). We provided a thorough explanation of the significance of the ASAG and its various uses. The sentence embeddings are generated from all four specified transfer learning models for both the desired and student replies in the dataset. The encoding of answers is contingent upon the words contained inside the answers, regardless of their sequential arrangement. The cosine similarity feature was computed for each student response and the corresponding desired answer. The feature was trained using three different regression methods: Polynomial, Linear, and Lasso regression. The USE model exhibited superior performance compared to other transfer learning models in the given task, with RMSE score of 0.997 and a Pearson correlation coefficient of 0.561. In this study, the performance of USE was compared to those of standard word embeddings, including Word2Vec, GloVe, and BERT. It is worth noting that no preprocessing or multiple feature training was employed during the evaluation procedure. The performance of USE was comparatively superior to that of other transfer learning models. The other transfer learning models have demonstrated inferior performance on the Mohler dataset in comparison to the conventional word embeddings. It was also determined that the utilization of USE can attain outcomes that are comparable to the most advanced techniques available without the inclusion of additional domain-specific data or the implementation of rigorous data pre-processing techniques.

FUTURE WORK

Despite the notable accomplishments attained through the utilization of USE without any preprocessing, multiple feature extraction, or training, there remains ample opportunity to further expand upon this research. Firstly, it is crucial to take into account the issue of demoting and eliminating stop words, as discussed in the works of [6], [20], and [21]. The sentence embeddings can be significantly impacted by the word alignment technique, as demonstrated by [20], due to the removal of inconsequential words. Furthermore, it is imperative to investigate several approaches for assigning sentence embeddings, such as sum of word embeddings (SOWE), mean-pooling and max-pooling, in addition to the conventional average of vectors.

REFRENCES

[1] B. Magnini, P. Rodríguez, D. Perez, A. Gliozzo, E. Alfonseca, and C. Strapparava, “About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment,” Revista signos: estudios de lingüística, ISSN 0035-0451, No. 59, 2005, pags. 325-343, Jan. 2005.

[2] O. Bukai, R. Pokorny, and J. Haynes, “An automated short-free-text scoring system: development and assessment,” in Proceedings of the Twentieth Interservice/Industry Training, Simulation, and Education Conference, 2006, pp. 1–11.

[3] C. Gütl, “e-Examiner: Towards a Fully-Automatic Knowledge Assessment Tool applicable in Adaptive E-Learning Systems,” 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:1850300

[4] S. Bailey and D. Meurers, “Diagnosing meaning errors in short answers to reading comprehension questions,” in Proceedings of the third workshop on innovative use of NLP for building educational applications, 2008, pp. 107–115.

[5] W.-J. Hou and J.-H. Tsao, “Automatic Assessment of Students’ free-Text Answers with Different Levels,” International Journal on Artificial Intelligence Tools, vol. 20, no. 02, pp. 327–347, 2011.

[6] M. Mohler, 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, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds., Portland,

Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 752–762. [Online]. Available: https://aclanthology.org/P11-1076

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[8] A. Vaswani et al., "Attention is all you need," Adv Neural Inf Process Syst, vol. 30, 2017.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[10] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[12] D. Cer et al., "Universal sentence encoder," arXiv preprint arXiv:1803.11175, 2018.

[13] J. Burstein, S. Wolff, and C. Lu, "Using Lexical Semantic Techniques to Classify Free-Responses," in Breadth and Depth of Semantic Lexicons, E. Viegas, Ed., Dordrecht: Springer Netherlands, 1999, pp. 227–244. doi: 10.1007/978-94-017-0952-1_11.

[14] D. H. Callear, J. Jerrams-Smith, and V. Soh, "CAA of short non-MCQ answers," 2001.

[15] H.-C. Wang, C.-Y. Chang, and T.-Y. Li, "Assessing creative problem-solving with automated text grading," Comput Educ, vol. 51, no. 4, pp. 1450–1466, 2008, doi: https://doi.org/10.1016/j.compedu.2008.01.006.

[16] T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge, "Towards robust computerised marking of free-text responses," Dec. 2002.

[17] L. F. Bachman et al., "A Reliable Approach to Automatic Assessment of Short Answer Free Responses," in Proceedings of the 19th International Conference on Computational Linguistics - Volume 2, in COLING '02. USA: Association for Computational Linguistics, 2002, pp. 1–4. doi: 10.3115/1071884.1071907.

[18] P. Thomas, "The Evaluation of Electronic Marking of Examinations," in Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education, in ITiCSE '03. New York, NY, USA: Association for Computing Machinery, 2003, pp. 50–54. doi: 10.1145/961511.961528.

[19] M. Mohler 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), A. Lascarides, C. Gardent, and J. Nivre, Eds., Athens, Greece: Association for Computational Linguistics, Mar. 2009, pp. 567–575. [Online]. Available: https://aclanthology.org/E09-1065.

[20] M. A. Sultan, C. Salazar, and T. Sumner, "Fast and Easy Short Answer Grading with High Accuracy," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, K. Knight, A. Nenkova, and O. Rambow, Eds., San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1070–1075. doi: 10.18653/v1/N16-1123.

[21] T. D. Metzler, P. G. Plöger, and G. Kraetzschmar, "Computer-assisted grading of short answers using word embeddings and keyphrase extraction," Master's thesis, 2019.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Adv Neural Inf Process Syst, vol. 26, 2013.

[23] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Trans Assoc Comput Linguist, vol. 5, pp. 135–146, 2017.

[24] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 670–680. doi: 10.18653/v1/D17-1070.

[25] M. E. Peters et al., "Deep Contextualized Word Representations," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. doi: 10.18653/v1/N18-1202.

[26] A. Radford and K. Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160025533