Routledge
Taylor & Francis Group

Interactive Learning Environments

ISSN: 1049-4820 (Print) 1744-5191 (Online) Journal homepage: https://www.tandfonline.com/loi/nile20

An automatic short-answer grading model for semi-open-ended questions

Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu & Fuzhen Zhuang

To cite this article: Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu & Fuzhen Zhuang (2019): An automatic short-answer grading model for semi-open-ended questions, Interactive Learning Environments, DOI: 10.1080/10494820.2019.1648300

To link to this article: https://doi.org/10.1080/10494820.2019.1648300

Published online: 30 Jul 2019.

INTERACTIVE LEARNING ENVIRONMENTS
https://doi.org/10.1080/10494820.2019.1648300

Routledge
Taylor & Francis Group

An automatic short-answer grading model for semi-open-ended questions

Lishan Zhang [b], Yuwei Huang[c], Xi Yang[c], Shengquan Yu[a] and Fuzhen Zhuang[d]

[a]Advanced Innovation Center for Future Education, Beijing Normal University, Beijing, People's Republic of China; [b]National Engineering Research Center for E-learning, Central China Normal University, Wuhan, People's Republic of China; [c]17Zuoye, Beijing, China; [d]Chinese Academy of Science, Institute of Computing Technology, Beijing, People's Republic of China

ABSTRACT

Automatic short-answer grading has been studied for more than a decade. The technique has been used for implementing auto assessment as well as building the assessor module for intelligent tutoring systems. Many early works automatically grade mainly based on the similarity between a student answer and the reference answer to the question. This method performs well for closed-ended questions that have single or very limited numbers of correct answers. However, some short-answer questions ask students to express their own thoughts based on various facts; hence, they have no reference answers. Such questions are called semi-open-ended short-answer questions. Questions of this type often appear in reading comprehension assessments. In this paper, we developed an automatic semi-open-ended short-answer grading model that integrates both domain-general and domain-specific information. The model also utilizes a long-short-term-memory recurrent neural network to learn the representation in the classifier so that word sequence information is considered. In experiments on 7 reading comprehension questions and over 16,000 short-answer samples, our proposed automatic grading model demonstrates its advantage over existing models.

ARTICLE HISTORY
Received 18 July 2019
Accepted 23 July 2019

KEYWORDS
Machine learning; short-answer grading; text analysis; classification; auto-grading

1. Introduction

Reading comprehension questions are commonly used to assess students' understanding of reading materials. Multiple-choice and short-answer reading comprehension questions are two typical types. Multiple-choice questions ask students to recognize the correct answer among several alternatives, while short-answer questions ask students to write answers in several phases or sentences. Generating good alternatives for multiple choice questions is non-trivial (Zhang & VanLehn, 2019). According to Chi's ICAP framework (Chi & Wylie, 2014), practicing with short-answer questions should also provide more benefits for students' learning than with multiple-choice questions. However, short answers cannot be autograded directly and manual grading is time-consuming and labor-intensive. Natural language processing (NLP) technology has been extensively applied to the autograding issue. Different from essay grading, which also requires dealing with NLP issues, the length of a short answer ranged from one phrase to one paragraph and the grading criteria focus on the content instead of the style (Higgins et al., 2004; Leacock & Chodorow, 2003; Mohler & Mihalcea, 2009; Pulman & Sukkarieh, 2005).

CONTACT Shengquan Yu yusq@bnu.edu.cn

Typical short-answer questions are closed-ended, namely, the correct answers are expected to match only several facts, which are also called as expectations, that are specified by the question authors. Short-answer questions of this type require students to explain objective facts without expressing their subjective opinions, such as the question below from Zhang, Shah, and Chi (2016).

Question: Why are there no potential energies involved in this problem?

Correct answer: Because the rock is the only object in the system, there are no potential energies involved.

However, some short-answer questions require students to express their subjective opinions based on a specified context. In this case, it is impossible to list all the expectations of the questions in advance. Hence, student answers must be graded based on a general grading rubric. On the other hand, questions of this type are not completely open-ended either, because students are expected to list specific facts and express their subjective opinions based on the facts. Therefore, we treat these short-answer questions as semi-open-ended questions. Many reading comprehension short-answer questions are semi-open-ended. The main goal of this paper is to propose and evaluate a model that can automatically grade these semi-open-ended short-answer questions. A semi-open-ended short-answer sample question is given below. It is a reading comprehension question with the reading material omitted. The question was originally written in Chinese, but translated into English.

Question: Based on the sentence "his lasting appeal of yo-heave-ho has become a kind of cultural landscape of Gumiao town" and your life experience, please explain your under-standing about the "yo-heave-ho" of King of sweet ferment rice.

Sample of correct answer: I think his yo-heave-ho shows diligent and intangible enthusiasm among people. In our life, vendors' shouting along street and their insistence on trading no matter how the weather is and what the season can make people feel warm-hearted in everyday life is.

The first sentence of the sample answer explains the meaning of "yo-heave-ho" in the article, which reflects warm-heated relationships among neighbors. The second sentence refers to an example in the student's personal life that also reflects a relationship of this type. To answer this question correctly, a student must accurately explain the meaning of "yo-heave-ho" and provide a similar example from her personal experience. The first part of a correct answer is highly predictable, but the latter part is flexible. This again explains why this type of question is called semi-open-ended. This characteristic makes autograding difficult.

The remainder of this article is organized as follows: We first review the purpose of automatic grading and its common technologies. Then, we introduce our automatic grading model and evaluate the model with 7 reading comprehension questions and over 16,000 labeled student answers. Last, we present the discussion and our conclusions.

2. Literature review

2.1. Purpose of automatic grading

There are two general purposes for implementing automatic grading models: for large-scale assessment and as part of the assessor in an intelligent tutoring system. The Educational Testing Service (ETS) is an industry initiator that pushes forward automatic grading of both short answers (c-rater) and essays (e-rater) (Attali & Burstein, 2004; Leacock & Chodorow, 2003). The ultimate goal of these raters is to accurately classify all student answers into the appropriate categories according to a grading rubric. In contrast, academic researchers are more interested in implementing automatic grading models for instructional purposes and adaptive learning (Luo & Litman, 2016; Zhang & VanLehn, 2017). For example, automatic grading models can be used to build conversational intelligent tutoring systems like AutoTutor, which is driven by expectation and misconception tailored (EMT) dialogue (Graesser et al., 2016; Shi et al., 2018). In each conversational turn, the EMT dialogue requires an assessor to automatically classify student answers into a list of expectations and

misconceptions, so that the intelligent agent knows how to respond (Nye, Graesser, & Hu, 2014). The goal of the assessor is consistent with that of an automatic grader. They both tried to classify students' text answers into multiple classes, either grades or predefined expectations. Besides these two general purposes, the underlying text analysis techniques of automatic grading can also provide new potentials for intelligent tutoring (Liu et al., 2019; Psotka & Chen, 2019).

2.2. Similarity-based automatic grading

As claimed by Burrows, Gurevych, and Stein (2015), short-answer questions typically have one or several clear correct answers that can be used as the reference and automatic grading can be conducted by calculating the similarity between a student answer and the reference answer (Graesser et al., 2004). This grading method has been used in intelligent tutoring systems in many domains (Azevedo et al., 2012; Dzikovska, Farrow, & Moore, 2013; Mcnamara et al., 2007; Rus & Graesser, 2006). Latent semantic analysis and latent Dirichlet allocation (Blei, Ng, & Jordan, 2003) are typically used to build topic vectors that represent student answers and reference answers that quantify similarities. Of all the similarity calculations, the cosine distance is one of the most common measures. From the perspective of machine learning, this automatic grading method is similar to the k-nearest neighbor algorithm, except that only answers with positive labels (correct answers) are typically provided. Automatic grading algorithms of this type are called similarity-based grading methods in this paper. The largest advantage of this method is that it requires almost no training data set. Several sample correct answers are sufficient for a question. However, the grading algorithm often requires a much larger dataset to calibrate the similarity calculation between students' answers and reference answers. The main disadvantage is that a similarity-based grading method can only recognize the variations of sample reference answers; all other answers will be treated as incorrect answers. Hence, it is important to ask the question in an appropriate way to ensure that correct student answers are highly predictable.

2.3. Automatic grading via a machine learning approach

With many data sets of labeled short answers available, researchers have started to build automatic grading models by using statistical machine learning algorithms. The development of a grading model via a machine learning approach can be divided into two main steps: feature engineering and classification. Feature engineering is used to extract features from student answers. It essentially defines the information that a classification algorithm can consider in making decisions. Most classification algorithms treat student answers as a bag of words without considering the word sequence, such as linear regression (Madnani et al., 2013), support vector machine (SVM) (Hou et al., 2010), and deep belief network (DBN) (Zhang et al., 2016). In our proposed model, a long short-term memory recurrent neural network (LSTM) is used to consider word sequence information. Regardless of which classification algorithm is used, the reference answers become optional in these approaches. However, machine learning grading methods require a much bigger set of labeled student answers than similarity-based methods so that automatic grading models can be learned from the data.

Each of the existing machine learning grading methods follows one of two general approaches: (1) Learn a question general grading model that can grade various questions in the same domain (Bailey & Meurers, 2008; Gütl, 2008; Nielsen, Ward, & Martin, 2008; Zhang et al., 2016). (2) Learn a question specific grading model for each question (Luo & Litman, 2016; Madnani et al., 2013; Yang, Zhang, & Yu, 2017). These two approaches have led to different feature engineering strategies.

While question general grading models are being built, the extracted student answer features must be question general as well. Hence, the reference answer, which is responsible for providing question specific information, becomes an essential part of the features. Similarities or differences between student and reference answers are the most common features. Other typical question general features are the length of the student answers, question difficulty, and student competency.

Therefore, a machine learning question general grading model is similar to an extended version of a similarity-based grading model.

In contrast, a question specific grading model typically does not require a reference answer, but directly learns the grading mechanism from the labeled student answers. However, this method requires many labeled student answers to train each question specific grading model. Collecting sufficiently manually labeled student answers is a key issue. The size of the set of labeled student answers, which is also called training dataset, may hinder the autograding performance. Since our reading comprehension questions are semi-open-ended and, thus, have no reference answers, we must develop a question specific grading model that is inferred from labeled student answers.

3. Proposed model

In developing a question specific autograding model, the size of the training data set may impede model performance. However, a human can learn how to grade from only a limited number of labeled student answers. This is because a human has a large amount of domain-general knowledge and is able to quickly learn the grading criteria by adapting her domain-general knowledge to the question. This has inspired machine learning researchers to develop many transfer learning algorithms for integrating domain-general and domain-specific information to improve classification accuracy (Pan & Yang, 2010). In our proposed model, domain-general and domain-specific information are integrated in the process of feature engineering. The domain-general information is extracted from Wikipedia and the domain-specific information is the set of labeled student answers.

As discussed above, existing autograding models often neglect word order information in student answers. In our proposed model, LSTM is used to encode word sequence information while building the classifier.

By improving both the feature engineering and the classification algorithm, we aim at answering two research questions:

- Does the integration of domain-general and domain-specific information help improve automatic grading performance?
- Do classifiers that incorporate word sequence information, such as LSTM, outperform classical classifiers that do not consider word sequence information?

The framework of our model is illustrated in Figure 1. In the remainder of this section, we describe how we integrated domain-general and domain-specific information in our feature engineering (left part in Figure 1) and how LSTM was used in our classification algorithm (right part in Figure 1).

3.1. Word embedding in student answers as the features

As Zhang et al. (2016) claimed, the features of an automatic grading model can come from the student's competency, the question, and the answer. But student's competency is not always available, especially in the context of assessment. This type of feature is not used in our feature engineering approach.

Words are commonly represented as embedding vectors such that words with similar meanings are close to each other in the vector space. LSA and LDA are the most common methods for transforming words into topic vectors. Recently, a new algorithm, namely, the continuous bag-of-words model (CBOW), was developed, which is based on deep learning technology. CBOW outperforms LSA on preserving linear regularities among words (Mikolov et al., 2013; Zhila et al., 2013). This model is used in our feature engineering approach.

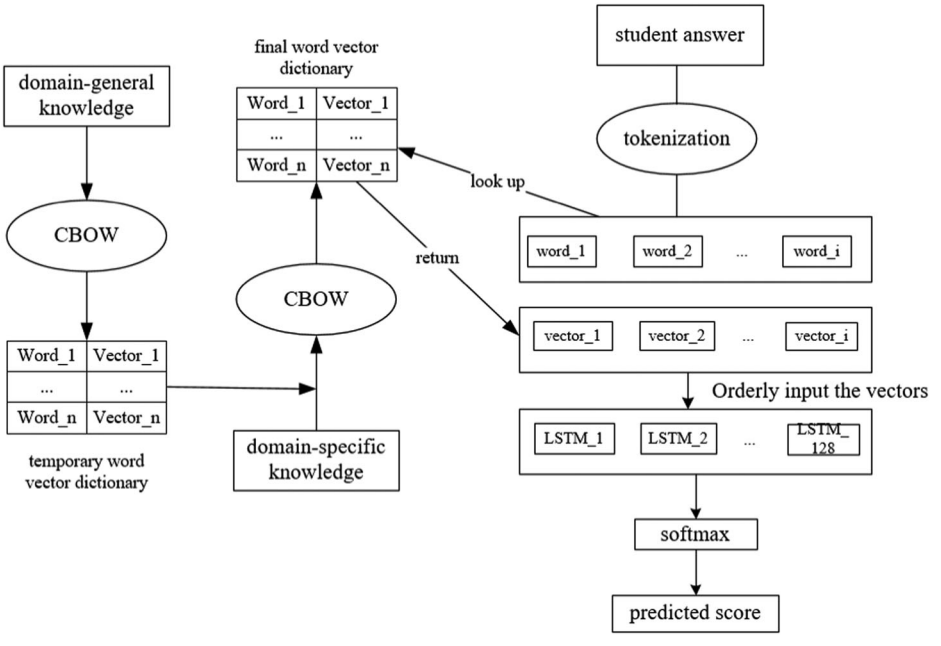

Figure 1. Automatic grading model overview.

3.1.1. CBOW model

The CBOW model essentially tries to predict the vector of each word based on its context words. It is a three-layered neural network that includes an input layer, projection layer and output layer. The model is illustrated in Figure 2. A brief description of the three layers is given below.

Input layer: For each target word w_k in a sentence, there are 2c context words, including c precedent words, which are denoted as $w_{k-c}, \ldots, w_{k-2}, w_{k-1}$, and c posterior words $w_{k+1}, w_{k+2}, \ldots, w_{k+c}$. Each word is initialized as a dimensional vector with a random value. The 2c context words are the input of the target word w_k. Their values are updated during the learning process.

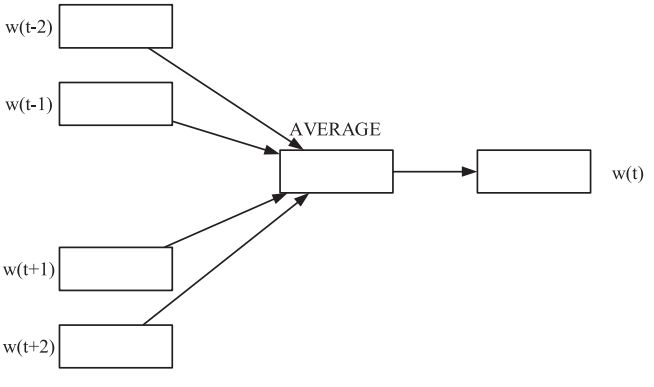

Figure 2. CBOW Model.

Projection layer: Calculate the average of all the vectors from the input layer. The average is treated as the vector of the context information of the target word w_k. The formula is shown as (1)

$$v_k = \frac{1}{2c} \sum_{i=k-c,\, i \neq k}^{k+c} v_i \tag{1}$$

where v_i is the vector of the context word w_i

Output layer: The predicted v_k should be similar to the actual v_k, but dissimilar to the vectors of other words, which form the negative set. The CBOW model tries to update the value of each target word to minimize the loss function in (2)

$$loss_{CBOW} = \log \prod_{k=1}^{|D|} \left\{ \sigma(v_k^T \theta^k) \prod_{j=1}^{|Neg_k|} [1 - \sigma(v_k^T \theta^j)] \right\} \tag{2}$$

where D is the entire word vocabulary of the corpus, Neg_k is a sample of the negative set, $\sigma(v_k^T \theta^k)$ is the likelihood that the predicted vector of the target word is its actual vector, and $\sigma(v_k^T \theta^j)$ is likelihood that the predicted vector is the vector of another word.

3.1.2. Integration of domain-general information with domain-specific information

Both domain-general information and domain-specific information are used to generate word vectors. Two training steps are implemented: In the first training step, the CBOW model is trained with Wikipedia data. At the beginning of the training, the word vector values are randomly initialized. After the corresponding loss function has been optimized, the second training step begins. Then, the CBOW model is trained with collected student answers. The main difference is that the initial word vector values are transferred from the result of the first step. Because the training data, namely, Wikipedia data, do not vary among questions, we only need to run the first training step once. The training data for the second step, namely, the set of student answers, vary among questions. Hence, the second training step must be run for each question to be autograded.

3.2. Classifier

We used LSTM (Gers, Schmidhuber, & Cummins, 1999) to learn the representation and softmax as our classifier to train the grading model. LSTM is a well-known recurrent neural network model in natural language processing. The largest advantage of this model is that it uses a memory node to consider the word sequence.

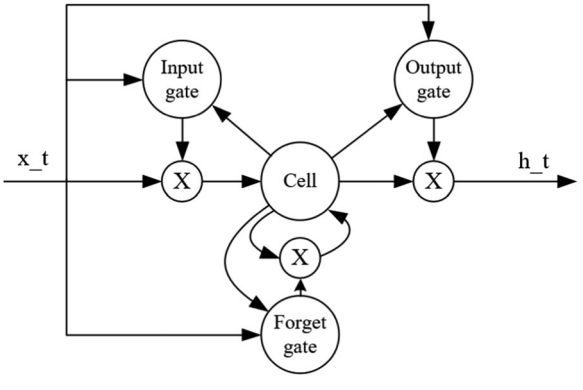

Figure 3. LSTM model.

The architecture of a unit in LSTM is illustrated in Figure 3. There are three gates, namely, the input, output, and forget gates, which control how the state is stored in the memory cell. Functions i_t, o_t, and f_t represent the activations of the input, output and forget gates, respectively, and c_t is the activation of the memory cell at step $t-1$, which is used to retain the information in the previous step. Vectors x_t and h_t are the input and output vectors, respectively, at step t. The activation functions are shown in (3). A softmax layer is used to combine the outputs of all the LSTM cells to predict the score of a student answer, as illustrated in Figure 3.

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= i_t \cdot \tan h(W_c x_t + U_c h_{t-1} + b_c) + f_t \cdot c_{t-1} \\ h_t &= o_t \cdot \tan h(c_t) \end{aligned} \tag{3}$$

where W_i, W_f, W_o, and W_c are the weight matrices of the input vector; U_i, U_f, U_o, and U_c are the weight matrices of the output vector in the previous step; and b_i, b_f, b_o, and b_c are bias vectors.

During the training phase, for each student answer, the classifier outputs the probability of each possible score; these probabilities compose vector s_i. The human-labeled score of the answer is represented by the corresponding vector s_i', in which the entry that corresponds to the true score is 1 and the remaining entries are 0. The classifier is trained with a cross-entropy loss function, as shown in (4). The Adam algorithm is used to optimize the loss function. After the classifier has been trained, the score with the highest probability is used as the score of a student answer to be graded.

$$C(s_i, s'_i) = -\frac{1}{n}\sum_{i=1}^{n}[s_i \ln s'_i + (1 - s_i)\ln(1 - s'_i)] \tag{4}$$

4. Evaluation

To evaluate our proposed model, we conducted experiments on 7 reading comprehension short-answer questions. Those included 5 questions in Chinese and 2 questions in English. Each question contained at least 2000 human graded student answers. In this section, we introduce the data sets and the measures and describe our experimental design and the results.

4.1. Datasets

The 5 reading comprehension questions in Chinese were obtained from the final exams of Grade 8 students. The set of these questions is called the Chinese Reading Comprehension Corpus (CRCC). The first two questions are generated based on two works of fiction that are required by the school syllabus. The remaining three are based on short reading materials that are given in the final exams. The 5 questions are listed in Table 1. We had two experienced teachers grade each

Table 1. Five Chinese reading comprehension questions.

ID	Question
CRCC 1	Please use your own words to write a recommendation for the fiction "Red Crag". The recommendation should help the readers grasp the gist of the fiction, and raise their interests
CRCC 2	"20,000 leagues under the sea" is a fiction. What is the most attractive point of this book to you?
CRCC 3	We have read "Mr. Bian meets Mr. Cai", where Mr. Bian failed to convince Mr. Cai. How do these two stories ("Mr. Bian meets Mr. Cai" and the reading material) inspire our interpersonal communication in the current age?
CRCC 4	Based on the sentence "his lasting appeal of yo-heave-ho has become a kind of cultural landscape of Gumiao town" and your life experience, please explain your understanding about the "yo-heave-ho" of King of sweet ferment rice
CRCC 5	Mr. Hwang adheres not to extend his business scale, and only sells 30 bottles of rice wine per day. Do you agree with him, and why?

student answer. For each question, they graded 50 student answers together to confirm the grading rubric and individually graded the remaining answers. They discussed with one another to resolve any conflicts. Quadratic-weighted kappa (QWKappa) is used to test their agreement. QWKappa is described in detail in the next section.

The 2 reading comprehension questions in English are from a Kaggle competition. The original Kaggle competition contained 10 questions. However, only two of them are reading comprehension questions. The kappa value is not available for this data set. The 2 questions are listed in Table 2. Table 3 lists the basic statistics of the 7 datasets.

4.2. Measures

QWKappa is used to evaluate the agreement among the grades that were assigned by different graders. We adopt QWKappa instead of the typical Cohen's Kappa because the former is able to capture the scores' order information. For example, suppose that a question can have up to 3 levels of scores (0,1,2), the first-grader scores an answer as 0, the second-grader scores the same answer as 1, and the third-grader scores the answer as 2. Although both the second and the third-grader disagree with the first-grader, it is clear that the second-grader is more similar to the first-grader than the third grader. The typical Cohen's Kappa cannot capture this difference, whereas QWKappa can. The details of the calculation of QWKappa can be found on Kaggle's website.[1]

4.3. Experiment design

The main objective of the experiment is twofold:

- To test the effectiveness of integrating domain-general and domain-specific information in the grading model;
- To test the effectiveness of our proposed model compared to other existing automatic grading models.

We implemented 7 automatic grading models, including our proposed model. Four out of the 7 grading models extracted features from student answers based on a bag-of-words model. The 4 classifiers are Logistic Regression (LR), which was used in (Madnani et al., 2013); Naïve Bayes (NB), which was used in (Levy et al., 2013); Decision Tree (DT), which was used in (Jimenez, Becerra, & Gelbukh, 2013); and SVM, which was used in (Hou et al., 2010). All 4 classifiers must have a single multi-dimensional vector as their input because these traditional models are unable to take word sequence information into consideration. Each dimension of the input vector represents a word and the value of the dimension is the TF-IDF value of the corresponding word. Hence, the number of dimensions of the input vector is equal to the number of words in the training data set. Only domain-specific information, namely, student answers, is used to train the 4 traditional classifiers. In contrast, the remaining 3 grading models all extracted features based on the CBOW model and respectively use domain-general information only, which is denoted as CBOW-G; domain-specific information only, which is denoted as CBOW-S; and both domain-general and domain-specific information, which is denoted as CBOW-GS.

Table 2. Two English reading comprehension questions.

ID	Question
Kaggle 1	Explain how pandas in China are similar to koalas in Australia and how they both are different from pythons. Support your response with information from the article.
Kaggle 2	Explain the significance of the word "invasive" to the rest of the article. Support your response with information from the article.

Table 3. Overview of the datasets.

Problem ID	Number of samples	Score level	#Distinct words	QWKappa
CRCC 1	2579	0–4	1071	0.9847
CRCC 2	2571	0–2	1644	0.9723
CRCC 3	2382	0–3	618	0.9427
CRCC 4	2458	0–4	655	0.9733
CRCC 5	2538	0–3	768	0.8319
Kaggle 1	2297	0–2	675	N/A
Kaggle 2	2033	0–2	804	N/A

We performed the same pre-processing step before training for all 7 grading models. We used Jieba[2] to segment the words in the Chinese corpus and scikit-learn to tokenize the English corpus. LR, NB, DT and SVM were implemented with scikit-learn. CBOW-S, CBOW-G and CBOW-GS were implemented with Keras (Chollet, 2015).

QWKappa between the predicted grades and the human resolved grades is calculated for each question. The experiment is conducted with 5-fold cross-validation and the average QWKappa over the 5 folds is used to compare the 7 automatic grading models. We test the following 3 hypotheses:

- *Hypothesis 1*. CBOW-GS outperforms both CBOW-G and CBOW-S because the first model extracts both domain-general and domain-specific information as the features.
- *Hypothesis 2*. CBOW-S and CBOW-G outperform the existing 4 grading models because the former two models take word sequence into consideration.
- *Hypothesis 3*. CBOW-GS performs the best among all 7 grading models.

4.4. Results

The performances of the 7 classifiers are detailed in Table 4. Two-tailed paired *t*-Test was used to test the significance of the difference. However, the performance of a grader is not guaranteed to be normal distributed, and the result reported by Table 4 confirms the view. Therefore, the Wilcoxon signed-rank test was also conducted, which is an alternative to the paired *t*-Test when the population cannot be assumed to be normally distributed. Table 5 reports both results, because the paired *t*-Test is more widely used, and the Wilcoxon signed-rank test is more suitable.

H1: QWKappa of CBOW-GS is significantly higher than those of CBOW-S and CBOW-G. Hence, hypothesis 1 is confirmed. Integrating both domain specific and domain general information helps improve the autograding performance.

H2: CBOW-S did not perform well and was outperformed by two of the bag-of-words grading models. SVM is the best-performing bag-of-words grading model. It significantly outperforms all the others. Although the average QWKappa of CBOW-G exceeds that of SVM, the difference is not significant. Hence, hypothesis 2 is not confirmed. This is probably because the domain general corpus is too sparse. The learned word vectors must be adapted to the domain specific corpus.

Table 4. QWKappa values between the outputs of the seven classifiers and resolved teacher grades.

	LR	NB	DT	SVM	CBOW_S	CBOW_G	CBOW_GS
CRCC 1	0.3697	0.1970	0.2959	0.4015	0.2213	0.4431	**0.4520**
CRCC 2	0.3915	0.1729	0.2556	0.4254	0.3752	0.4825	**0.4983**
CRCC 3	0.7913	0.6340	0.8108	0.8680	0.7276	0.8364	**0.8694**
CRCC 4	0.5142	0.2954	0.4333	0.5789	0.5693	0.5612	**0.5911**
CRCC 5	0.6270	0.4465	0.6288	0.6522	0.4214	0.6754	**0.7058**
Kaggle 1	0.5604	0.5046	0.4558	0.5905	0.5947	0.6126	**0.6430**
Kaggle 2	0.5482	0.5644	0.4433	0.5695	0.5655	0.5717	**0.6103**
Mean	0.5432	0.4021	0.4748	0.5837	0.4964	0.5976	**0.6243**
(SD)	(0.1431)	(0.1819)	(0.1914)	(0.1549)	(0.1679)	(0.1307)	**(0.1378)**

Table 5. Pair-by-pair comparison.

	NB CBOW_S	DT CBOW_G	SVM CBOW_GS
LR	***t* = 4.223, *p* = 0.006**	*t* = 3.131, *p* = 0.020	***t* = −5.030, *p* = 0.002**
	***Z* = −2.197, *p* = 0.028**	*Z* = −1.859, *p* = 0.063	***Z* = −2.366, *p* = 0.018**
	t = 1.260, *p* = 0.254	***t* = −6.615, *p* = 0.001**	***t* = −16.163, *p* < 0.001**
	Z = −0.845, *p* = 0.398	***Z* = −2.366, *p* = 0.018**	***Z* = −2.366, *p* = 0.018**
NB		*t* = −1.663, *p* = 0.147	***t* = −4.882, *p* = 0.003**
		Z = −1.524, *p* = 0.128	***Z* = −2.366, *p* = 0.018**
	t = −2.281, *p* = 0.063	***t* = −4.977, *p* = 0.003**	***t* = −6.042, *p* = 0.001**
	Z = −1.859, *p* = 0.063	***Z* = −2.366, *p* = 0.018**	***Z* = −2.366, *p* = 0.018**
DT			***t* = −5.578, *p* = 0.001**
			***Z* = −2.366, *p* = 0.018**
	t = −0.406, *p* = 0.699	***t* = −4.763, *p* = 0.003**	***t* = −6.248, *p* = 0.001**
	Z = −0.676, *p* = 0.499	***Z* = −2.366, *p* = 0.018**	***Z* = −2.366, *p* = 0.018**
SVM	*t* = 2.417, *p* = 0.52	*t* = −1.159, *p* = 0.291	***t* = −4.267, *p* = 0.005**
	Z = −2.028, *p* = 0.043	*Z* = −1.183, *p* = 0.237	***Z* = −2.366, *p* = 0.018**
CBOW_S		***t* = −2.553, *p* = 0.043**	***t* = −3.389, *p* = 0.015**
		***Z* = −0.028, *p* = 0.043**	***Z* = −2.366, *p* = 0.018**
CBOW_G			***t* = −6.767, *p* = 0.001**
			***Z* = −2.371, *p* = 0.018**

Notes: The two-tailed paired *t*-Test and Wilcoxon signed-rank test are used to compare the results of the autograding models on the 7 questions.

H3: CBOW-GS significantly outperformed all the other grading models. Hypothesis 3 is confirmed.

The results demonstrate that both domain-general information and word sequence should be included in the grading model. LR, NB, DT and SVM all use a bag-of-words model in feature engineering. A bag-of-words model builds a dictionary of all the words, where each word is represented as a float value. Then, a student answer is represented as a vector of float values. The sequence of the words is lost. LSTM has the advantage of taking a sequence of word vectors as the input and word vectors can be generated from either domain-specific or domain-general knowledge, or both. Hence, we are able to build a higher-performing grading model than the existing ones by combining CBOW and LSTM.

CBOW-S does not perform well, probably because the domain-specific corpus is not large enough to train reliable word vectors. Indeed, researchers usually train word vectors with large-scale auxiliary data sets. This also explains why CBOW-G performs much better.

We take two questions as examples below to show the advantages of CBOW and LSTM.

Question 1. Based on the sentence "his lasting appeal of yo-heave-ho has become a kind of cultural landscape of Gumiao town" and your life experience, please explain your understanding about the "yo-heave-ho" of King of sweet ferment rice.

Answer 1.1 The "yo-heave-ho" of King of sweet ferment rice reflects that he is rustic and hardworking (勤劳). In my life, my mom calls me up every morning. Her voice is rustic as well. (correct)

Answer 1.2 The "yo-heave-ho" of King of sweet ferment rice represents his diligence (勤奋). This kind of diligence is worth to be inherited to our age of life (correct).

Question 2. We have read "Mr. Bian meets Mr. Cai", where Mr. Bian failed to convince Mr. Cai. How do these two stories ("Mr. Bian meets Mr. Cai" and the reading material) inspire our interpersonal communication in the current age?

Answer 2.1 The way of Mr. Bian[扁鹊]'s talking is too harsh. In contrast, Mr. Zou [邹忌] (the figure in the reading material) convince his king by analogy, instead of direct criticizing. Therefore, the king accepts Mr. Zou's suggestion. (correct)

Answer 2.2 The way of Mr. Zou[邹忌]'s talking is too harsh. In contrast, Mr. Bian[扁鹊] (the figure in the reading material) convince his king by analogy, instead of direct criticizing. Therefore, the king accepts Mr. Zou's suggestion. (incorrect)

Note: Both the questions and the answers were originally written in Chinese. The most important key words are underlined and their original Chinese key words are given in parenthesis.

The two answers to question 1 demonstrate the advantage of CBOW model when both domain-general and domain-specific information are being used. The bag-of-words based grading model can recognize answer 1.1 as a correct answer due to the existence of the keyword [勤劳], but failed in grading answer 1.2 because the word [勤劳] is replaced with [勤奋]. However, the CBOW model can help recognize that [勤劳] and [勤奋] have similar meaning and grade answer 1.2 correctly as well.

The two answers to question 2 show the advantage of LSTM when the word's sequence is taken into consideration. Answer 2.1 is correct but answer 2.2 is incorrect because [扁鹊] and [邹忌] are exchanged. Vectors that represent the two answers have no difference when a bag-of-words model is adopted. However, LSTM is able to capture the difference and grade them correctly.

5. Discussion

The automatic grading model has less information for making decisions when a student's answer is short. It is similar to a topic modeling algorithm that often performs worse when it is used for Tweet analysis compared to article analysis (Mehrotra et al., 2013; Xiang et al., 2012). Additionally, the single reference answer is absent when the question is relatively open. Hence, short answer grading for semi-open-ended questions is a difficult problem. By integrating domain-general and domain-specific information, we significantly improved the performance of our automatic short-answer grading model. Although we only run experiments on Chinese and English reading comprehension short-answer questions, the proposed model can be potentially extended to any short-answer grading task that has no reference answer and many graded student answers.

Our proposed grading model performed differently on the 5 Chinese short-answer questions. The grading model performed poorly on the first two questions, but much better on the remaining three. This is probably because the first two questions asked students to write their answers based on an entire book, whereas the remaining three were based on several paragraphs of text. As a result, the student answers to the first two questions were more diverse. The diversity of the answers increased the grading difficulty. According to Table 3, the numbers of distinct words in the first two questions exceed those of the remaining questions. This is consistent with our conjecture. These differences also reflect the specificity of semi-open-ended short-answer questions compared to closed-ended short-answer questions. Closed-ended short-answer questions elicit specific facts; hence, substantial diversity among the answers is not expected, regardless of whether the answers are derived from an entire book or a short text. For example, a question in computer literacy may ask students to explain how an operating system loads a text file from hard disk to memory. The question covers multiple chapters in a textbook. However, the expectations of the question are specific; hence, the answers are closed-ended. In contrast, students have more freedom to express their own thoughts in answering semi-open-ended questions. Therefore, the diversity of the answers increases with the scope of the question on which they are based. In this case, the corresponding automatic grading model requires more labeled data to capture the diversity.

The largest disadvantage of our proposed model is the requirement of many graded student answers for each target question. As a result, the grading model may be only useful for large-scale applications that may have millons of student answers to a limited number of questions. For example, an instructor should consider similarity-based grading methods instead of our proposed model if she wants to autograde only one class of students. However, intelligent tutoring system

or learning assistant system designers should consider integrating our proposed model into their system as an assessor (Sottilare et al., 2018; Vanlehn, 2006), if the tutoring system is going to benefit hundreds of thousands of students. For example, the WISE project, which has been used in many countries, recently implemented an English short-answer grader in the system so that instructors can immediately assess a student's progress (Tansomboon et al., 2017). In addition, our proposed grading model can be potentially used in large-scale assessment. However, according to our reported measurements, human raters still significantly outperformed our grading models. The grading model must be further improved to fulfill the requirements of a prestigious assessment. Collecting more data may help further improve the performance.

6. Conclusions and future work

We defined semi-open-ended short-answer questions that ask students to express their subjective thoughts based on some facts in their answers. To build an automatic grading model for semi-open-ended short-answer questions, we integrated domain-general information from Wikipedia and domain-specific knowledge from the graded student short answers with the help of CBOW and generated word vectors to feed our LSTM-based classifier. Our experiments demonstrated that integrating both domain-general and domain-specific information significantly improved the automatic grading performance on semi-open-ended questions. Taking advantage of word sequence information by using LSTM improved the grading accuracy as well.

As we discussed, training the grading model requires many graded student answers. More answers are needed when the corresponding semi-open-ended questions are based on a book rather than a short text. Since no grading rubric is currently used in our grading model, in the next step, we will explore the integration of information from a grading rubric to further improve the performance of the autograding model.

Notes

1. https://www.kaggle.com/c/asap-sas#evaluation.
2. https://github.com/fxsjy/jieba.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Natural Science Foundation of China: [grant number 61807004] and Philosophy and Social Sciences Research of the Chinese Ministry of Education: [grant number 16JZD043].

Notes on contributors

Lishan Zhang is an associate professor at Central China Normal University. He received a PhD in Computer Science from Arizona State University. He has published over 20 peer-reviewed academic papers. His research interests include intelligent tutoring systems, student modeling for personalized learning and educational data mining.

Yuwei Huang is an NLP engineer in Sunny Education Inc. He received the BS degree from Beijing University of Chemical Technology in 2018.

Xi Yang is an AI Researcher in Sunny Education Inc. She received the PhD degree from Peking University in 2013. She has published over 10 peer-reviewed academic papers. Her research interests include machine learning, natural language processing and artificial intelligence in education.

Shengquan Yu is a Professor at Beijing Normal University. He received a PhD in Educational Technology from Beijing Normal University. His research fields include mobile and ubiquitous learning, ICT and curriculum integration,

network learning technology, and education informatization policy. He has published about 100 peer-reviewed academic papers, four popular science books and three scholarly monographs.

Fuzhen Zhuang is an associate professor in the Institute of Computing Technology, Chinese Academy of Sciences. He received his PhD degree in Computer Software and Theory from Chinese Academy of Sciences. His research interests include transfer learning, machine learning, data mining, multi-task learning and recommendation systems. He has published more than 80 papers in some prestigious refereed journals and conference proceedings.

ORCID

Lishan Zhang http://orcid.org/0000-0003-0830-2399

References

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2.

Azevedo, R., Landis, R., Feyzi-Behnagh, R., Duffy, M., Trevors, G., Harley, J., … Hossain, F. (2012). *The Effectiveness of Pedagogical Agents' Prompting and Feedback in Facilitating Co-adapted Learning with MetaTutor. Intelligent Tutoring Systems. ITS 2012. Lecture Notes in Computer Science, vol 7315. Springer, Berlin, Heidelberg*. Springerhttps://doi.org/10.1007/978-3-642-30950-2_27.

Bailey, S., & Meurers, D. (2008). *Diagnosing meaning errors in short answers to reading comprehension questions*. The workshop on innovative use of Nlp for building educational applications.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Chollet, F. (2015). *Keras: Deep learning library for theano and tensorflow*, vol. 7: p. 8. Retrieved from https://keras.io/k

Dzikovska, M. O., Farrow, E., & Moore, J. D. (2013). Combining semantic interpretation and statistical classification for improved explanation processing in a tutorial dialogue system. *International Conference on Artificial Intelligence in Education*, *7926*, 279–288.

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. *Artificial Neural Networks. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, *2*, 850–855.

Graesser, A. C., et al. (2004). Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 180–192.

Graesser, A. C., Cai, Z., Baer, W. O., Olney, A. M., Hu, X., Reed, M., & Greenberg, D. (2016). Reading comprehension lessons in AutoTutor for the center for the study of adult literacy. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 288–293). New York, NY: Taylor & Francis Routledge.

Gütl, C. (2008). Moving towards a fully automatic knowledge assessment tool. *International Journal of Emerging Technologies in Learning (Ijet)*, *3*(1), 36–44.

Higgins, D., et al. (2004). *Evaluating multiple aspects of coherence in student essays*. Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL, pp. 185–192.

Hou, W., et al. (2010). *Automatic assessment of students' free-text answers with support vector machines*. IEA/AIE'10 proceedings of the 23rd International conference on industrial engineering and other applications of applied intelligent systems - volume part I, pp. 235–243.

Jimenez, S., Becerra, C., & Gelbukh, A. (2013). *Softcardinality: Hierarchical text overlap for student response analysis*. Second joint conference on lexical and computational semantics (*SEM), volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), pp. 280–284.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405.

Levy, O., et al. (2013). UKP-BIU: Similarity and entailment metrics for student response analysis. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, *2*, 285–289.

Linn, M. C., & Slotta, J. D. (2010). *Designing science instruction using the web-based inquiry science environment (WISE)*.

Liu, Z., Yang, C., Rüdian, S., Liu, S., Zhao, L., & Wang, T. (2019). Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums. *Interactive Learning Environments*, *27*, 598–627. doi:10.1080/10494820.2019.1610449

Luo, W., & Litman, D. J. (2016). *Determining the quality of a student reflective response*. FLAIRS.

Madnani, N., et al. (2013). *Automated scoring of a summary-writing task designed to measure reading comprehension*. Proceedings of the eighth workshop on innovative use of NLP for building educational applications, pp. 163–168.

Mcnamara, D. S., et al. (2007). *iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies*.
Mehrotra, R., et al. (2013). *Improving LDA topic models for microblogs via tweet pooling and automatic labeling*. Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval, pp. 889–892.
Mikolov, T., et al. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.
Mohler, M., & Mihalcea, R. (2009). *Text-to-text semantic similarity for automatic short answer grading*. Conference of the European chapter of the association for computational linguistics.
Nielsen, R. D., Ward, W., & Martin, J. H. (2008). *Learning to assess low-level conceptual understanding*. FLAIRS.
Nye, B. D., Graesser, A. C., & Hu, X. (2014). Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, *24*(4), 427–469.
Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.
Psotka, J., & Chen, N. (2019). The new potentials for intelligent tutoring with learning analytics approaches. *Interactive Learning Environments*, *27*, 583–584. doi:10.1080/10494820.2019.1612888
Pulman, S. G., & Sukkarieh, J. Z. (2005). *Automatic short answer marking*. In Proceedings of the second workshop on Building Educational Applications Using NLP (pp. 9-16). Association for Computational Linguistics.
Rus, V., & Graesser, A. C. (2006, July 16–20). *Deeper natural language processing for evaluating student answers in intelligent tutoring systems*. National conference on artificial intelligence and the eighteenth innovative applications of artificial intelligence conference, Boston, MA.
Shi, G., Lippert, A., Shubeck, K. T., Fang, Y., Chen, S., Pavlik, P. I., … Graesser, A. C. (2018). Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension. *Behaviormetrika*, *45*(2), 615–633.
Sottilare, R. A., Baker, R., Graesser, A., & Lester, J. (2018). Special issue on the generalized intelligent framework for tutoring (GIFT): Creating a stable and flexible platform for innovations in AIED research. *International Journal of Artificial Intelligence in Education*, *28*(2), 131–151.
Tansomboon, CFM, Gerard, L, & Vitale, J. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757.
Vanlehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.
Xiang, G., et al. (2012). *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus*. Proceedings of the 21st ACM international conference on information and knowledge management, pp. 1980–1984.
Yang, X., Zhang, L., & Yu, S. (2017). *Can short answers to open response questions be auto-graded without a grading rubric?*. International Conference on Artificial Intelligence in Education 594-597.
Zhang, Y., Shah, R., & Chi, M. (2016). *Deep learning + student modeling + clustering: A recipe for effective automatic short answer grading*. International Educational Data Mining Society 562-567.
Zhang, L, & VanLehn, K. (2019). Evaluation of auto-generated distractors in multiple choice questions from a semantic network. *Interactive Learning Environments*. DOI: 10.1080/10494820.2019.1619586
Zhang, L., & VanLehn, K. (2017). Adaptively selecting biology questions generated from a semantic network. *Interactive Learning Environments*, *25*(7), 828–846.
Zhila, A., et al. (2013). *Combining heterogeneous models for measuring relational similarity*. Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 1000–1009.