Content-Based Methods in Peer Assessment of Open-Response Questions in MOOCs to Grade Students as Authors and as Graders

Oscar Luaces[a,*], Jorge Díez[a], Amparo Alonso-Betanzos[b], Alicia Troncoso[c], Antonio Bahamonde[a]

[a] *Artificial Intelligence Center*
University of Oviedo
33204 Gijón, Spain
[b] *Dept. of Computer Science*
Faculty of Informatics
University of A Coruña
15071 A Coruña, Spain
[c] *Dept. of Computer Science*
Pablo de Olavide University
41013 Sevilla, Spain

Abstract

Massive Open Online Courses (MOOCs) use different types of assignments in order to evaluate student knowledge. Multiple-choice tests are particularly apt given the possibility for automatic assessment of large numbers of assignments. However, certain skills require open responses that cannot be assessed automatically yet their evaluation by instructors or teaching assistants is unfeasible given the large number of students. A potentially effective solution is peer assessment whereby students grade the answers of other students. However, to avoid bias due to inexperience, such grades must be filtered. We describe a factorization approach to grading, as a scalable method capable of dealing with very high volumes of data. Our method is also capable of representing open-response content using a vector space model of the answers. Since reliable peer assessment requires students to make coherent assessments, students can be motivated by their assessments reflecting not only their own answers but also their efforts as graders. The method described is able to tackle both these aspects simultaneously. Finally, for a real-world university setting in Spain, we compared grades obtained by our method and grades awarded by university instructors, with results indicating a notable improvement from using a content-based approach. There was no evidence that instructor grading would have led to more accurate grading outcomes than the assessment produced by our models.

[*] Corresponding author: Tel: +34 985 182 028
Email addresses: oluaces@uniovi.es (Oscar Luaces), jdiez@uniovi.es (Jorge Díez), ciamparo@udc.es (Amparo Alonso-Betanzos), atrolor@upo.es (Alicia Troncoso), abahamonde@uniovi.es (Antonio Bahamonde)

Keywords: Peer Assessment, Factorization, Preference Learning, Grading Graders, MOOCs

1. Introduction

Massive open online courses (MOOCs) offer promising new educational opportunities and have focused the attention of many researches in terms of improving the educational experience of students. Since MOOCs attract thousands of students, assessment in particular – in order to provide feedback to students and to guarantee the quality of qualifications – is a problematic issue, since the vast numbers of students enrolled implies a huge or even impossible burden for instructors and teaching assistants. Assessment is therefore one of the most complex and challenging applications of big data in education.

We tackled the challenge of evaluating open-response questions, adopting, as our basic strategy, peer assessment [1, 2, 3, 4, 5, 6, 7, 8], whereby students evaluate the anonymized answers of other students participating in the same course. Students, in their role as graders, receive a set of detailed rules (called a rubric) designed to ensure uniform assessment. However, since students typically have no experience of assessing peers, grading must deal with the effects of inconsistent and subjective evaluation. Yet peer assessment also has an important pedagogical function in that a deeper understanding of course content is obtained when students read and are required to assess other students' answers. The two main peer assessment streams are cardinal and ordinal.

In cardinal peer assessment, grades are numbers or categorical labels with straightforward numerical semantics. If we have a sufficiently large number of grades for each assignment, then the correct grade could be approximated by computing the mean or the median [2]. Means have been reported to be more consistently accurate with respect to the rubric than staff grades [3]. However, one problem with the cardinal approach is that students cannot be charged with the job of grading large numbers of answers and another issue is that it is affected by the lack of student experience in assessment.

In the ordinal approach to peer assessment, graders rank answers in terms of their quality [5, 8, 7] – clearly an easier task for inexpert graders than cardinal grading, as evidenced by the considerably higher reliability reported for ordinal compared to cardinal assessment [9, 10, 11, 1, 12]. (See [13] for an interesting discussion of cardinal and ordinal peer grading from a psychological point of view).

Another approach to assessment is content-based methods, which use information retrieval techniques, for instance, a preference approach to learning the relevance of documents [14]. These methods require some shallow linguistic processing and also frequently require assistance from the instructor. Methods include comparing several ideal answers (references) with student answers or labelling a subset of answers with correct grades that are then extended to the whole set of answers using a machine learning algorithm.

As far as we are aware, no existing peer-assessment method takes into account the content of student responses to open questions. Since peer-assessment methods function like collaborative filters that recommend a grade for each answer, their predictive power could be enhanced using available information about answers.

We describe an approach that combines the strengths of ordinal collaborative filters and content-based recommenders. We use a factorization method to train a utility function that estimates consensus in rankings of answers. This approach – inspired by a preference learning framework [15, 9] – was used in previous research by us [6, 7, 16, 17]. Answers can be represented by vectors of features, which have been acknowledged to be crucial for the success of peer assessment [18, 19]. If no other information is available, features only capture a binary identification of answers and graders, reflecting a pure collaborative approach. However, our factorization method allows representations that include other information about the answers. Unlike other approaches, our proposed method does not need any self-grading of answers or any previous grading by instructors.

We also propose a method to grade students as graders, as student grading is a potentially powerful motivational aspect in learning. This would require announcing, before starting a course, that students' final grades would be calculated as a linear combination of answers both authored by and graded by students.

Below we formally describe our assessment method and results for a real-world dataset based on a computer science assignment issued to students at three Spanish universities, reporting discrepancies for our methods with instructor grades that were similar or lower than discrepancies between instructors. We tested both collaborative filtering and content-based representations, finding that the latter achieved considerably better results. Our proposal for grading graders also obtained good results, which improved, furthermore, in line with the number of answers evaluated by graders. The use of content also improved scores in most cases.

The pilot experience described here illustrates the viability of our proposed method, which is entirely scalable to a Big Data scenario. Both the dataset and the algorithms used here will be made publicly available on publication of the paper.

2. Related works

As with recommender systems, automatic assessment methods can be split into two groups: those that use answer content provided by students and those that function as collaborative filters.

Some interesting content-free assessments have been described [3, 5] with authors emphasizing the importance of assessing grader accuracy. In fact, accurate evaluations are crucial to obtaining reliable data so one way to encourage good-quality grading is to include grading of peer-grading assessments as part of the student's final grade.

Shah et al. [19] propose using methods that include some kind of *dimensionality reduction*, e.g., clustering, and using features to represent the issues involved in assessment. Although their proposals are very abstract, the factorization method proposed here offers a suitable framework for implementing both approaches.

In the area of content-based systems, the most widely used option is to combine shallow Natural Language Processing with Machine Learning, that is, methods borrowed from the Information Retrieval field. Broadly speaking, we can distinguish between matching and categorization methods.

Matching methods compare students' answers against some reference (ideal answer) or template; Pérez-Marín et al. [20] made a detailed survey of published algorithms which used this paradigm. Rodrigues and Oliveira [21] matched students' answers with references by computed cosine similarity after preprocessing. Both references and students' answers were represented using the vector space model (VSM) [22], in which each word is the index of a vector whose values – which may be weighted using different strategies – record the presence or frequency of a word in a document.

To deal with answer content, some authors have used matching methods that exploit coincidences between groups of words, with the aim being to take into account the syntactic structure of documents without penalizing the process with a deep analysis. A key tool in this case is a metric of document similarity called BLEU [23], devised to assess the quality of machine translations. Given a set of reference translations, BLEU computes scores for candidate translations based on the co-occurrence of n-grams in the references and candidate translations. A modified version of BLEU was used by Noorbehbahani and Kardan [24] to build a system for automatic assessment of open-ended answers.

The main disadvantage of content-based methods is that they do not consider synonyms. Since we cannot reasonably expect students to use exactly the same words as used in reference answers, a certain degree of semantic analysis is necessary to fairly compare students' answers with references. One way to overcome this problem is to use *Latent Semantic Analysis* (LSA) [25], which projects the matrix of VSM representations of all answers (usually called the *term-document* matrix) into a smaller dimensional space using the singular value decomposition (SVD) of the matrix. This robust information retrieval method thus manages to capture the implicit semantics of a set of documents.

A pilot LSA study that evaluated six students' answers to three questions in the computer science domain reported high precision despite a small dataset [26]. LSA was also used to assess participants in a professional development program according to five attitudinal categories of free-form text responses [27], with preprocessing – based on standardization, stop-word removal and Porter stemming – implemented in order to obtain the term-document matrix. Pérez et al. [28] proposed combining BLEU and LSA to assess open-ended answers. In our factorization approach – a generalization of the SVD matrix decomposition method – the decomposition aims to optimize a loss function and so improve predicted outcomes.

Rodrigues and Oliveira [21], mentioned above, included semantic analysis in their proposed cosine similarity method, whereby two words were considered to be similar if they were related in the WordNet semantic network.

Another content-based approach is an adaptation of text categorization, whereby a reduced set of answers is graded by the instructor and then processed by an ordinal classifier that learns from the supervised dataset of answers – represented by a feature vector – and grades. Useful for this purpose is an algorithm called *CarmelTC* that implements a naïve Bayes classifier [29]. There is also an approach [30] using a support vector machine (SVM) to evaluate creative problem-solving from open-ended responses. A comparison between the results obtained by LSA, SVM and a standard regression method showed that correlation with instructor scores was highest for SVM.

As for the task of grading graders, a probabilistic approach has been described for modeling reviewers' handling of scientific papers [31], based on the use of an expectation maximization (EM) algorithm to estimate means and variances. The goal was to improve assessment quality, but implicitly the method produced a ranking of reviewers according to bias. Kurokawa et al. [32] described an evaluation of research proposals that incorporated a correction for peer bias. This study was motivated by a US National Science Foundation pilot experience in which principal investigators peer-reviewed each other's proposals. Since this process could be affected by serious ethical issues, it was necessary to somehow correct for peer bias, which was done by using features to select specific projects to be assigned to each reviewer. Content, again, was crucial to this peer-assessment method.

Raman and Joachims [5, 8] described several probabilistic methods used to grade graders as well as rank answers; our approach is similar, with the main difference being the grader grading method and the evaluation of the proposal.

This research is an extension of research described in two conference papers [17, 16], but now with a considerably broader experimental setting. Moreover, the performance of our method has been considerably enhanced over that described in previous research that did not use answer content [7].

3. Learning method

Consider a set of students who receive an assignment to complete in the form of answers submitted in text documents. After submitting their answers, the students are asked to play the role of graders of a set of answers (we will refer to these student-graders simply as graders). The purpose is to build a function that can assess answers according to the grades awarded by these graders.

Formally, let $\mathcal{G}$ be a set of *graders* and let $\mathcal{A}$ be a set of *answers*. Each grader g has received a subset $\mathcal{A}_g \subset \mathcal{A}$ of answers to evaluate. The initial data to infer a grading function is the *assessment matrix*, $\boldsymbol{M}$, which contains the scores given by the student-graders:

$$\boldsymbol{M}(\boldsymbol{g}, \boldsymbol{a}) \in [0, 10], \tag{1}$$

where $\boldsymbol{g} \in \mathcal{G}$, and $\boldsymbol{a} \in \mathcal{A}_g \subset \mathcal{A}$.

Both graders and answers are represented by vectors of features (we will use the same symbols for these and their vectorial representations). In the simplest case, a grader (answer) can be identified by a vector of binary values with all zeros but a single 1 in the component indexed by itself in $\mathcal{G}$ (respectively $\mathcal{A}$). This elementary representation can be enriched with features describing additional aspects of graders or answers. In Section 4 we illustrate the positive effect of this enriched representation for the answers.

The goal of any peer-assignment method is to obtain an absolute ranking of answers from the scores in $\boldsymbol{M}$. Note that this goal is essentially the same as in *recommender systems*. We will use an approach already used with success in this field.

Firstly, however, it should be noted that a straightforward way to rank answers is to compute the mean grade for each answer and order answers accordingly. Nonetheless, this approach is only accurate when each answer has been assessed by a large number of graders, which is not the case in a peer-assessment context. In general, matrix $\boldsymbol{M}$ is going to be very sparse, as only a few answers will be graded by each grader.

We trained a scoring function to fill the matrix $\boldsymbol{M}$ and obtained the final ranking of the mean score for each answer for all the graders. This scoring function was induced based on a qualitative point of view. We used the *preference learning* approach to avoid grader subjectivity: in other words, we focused on the relative order of answers for each grader and not on the score values. We thus built a set of *preference judgements*, $\mathcal{D}$, given by triples of a grader $\boldsymbol{g}$ and a pair of answers $(\boldsymbol{a}_b, \boldsymbol{a}_w)$ in $\mathcal{A}_g$ such that

$$\boldsymbol{M}(\boldsymbol{g}, \boldsymbol{a}_b) > \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{a}_w) \Rightarrow [\boldsymbol{g}, \boldsymbol{a}_b, \boldsymbol{a}_w] \in \mathcal{D}. \tag{2}$$

Note that since answers with the same score do not yield any relative order, ties were discarded when generating the dataset of preference judgements.

The strategy followed to obtain the ranking started with a double *embedding*, that is, mapping both answers and graders into a common Euclidean space $\mathbb{R}^k$:

$$\mathbb{R}^{|\mathcal{G}|} \to \mathbb{R}^k, \quad \boldsymbol{g} \mapsto \boldsymbol{W}\boldsymbol{g}; \tag{3}$$

$$\mathbb{R}^{|rep(\mathcal{A})|} \to \mathbb{R}^k, \quad \boldsymbol{a} \mapsto \boldsymbol{V}\boldsymbol{a}. \tag{4}$$

Note that using an enriched vectorial representation of the answers in addition to the identifier of the answer itself means that the representation of answers ($rep(\mathcal{A})$) may have a higher dimension than the number of answers.

From dataset $\mathcal{D}$ and the embedding, we define the *individual grade* as a function of graders and answers as follows:

$$f(\boldsymbol{g}, \boldsymbol{a}) = \langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a} \rangle. \tag{5}$$

Since this function estimates the grade given by any grader $\boldsymbol{g}$ to any answer $\boldsymbol{a}$, it was used to complete the assessment matrix. We could thus compute the *final grade* for each answer as the mean of all its grades.

$$\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \mathrm{f}(\boldsymbol{g},\boldsymbol{a}) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a}\rangle = \\ \left\langle \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a}\right\rangle = \langle \boldsymbol{W}\bar{\boldsymbol{g}}, \boldsymbol{V}\boldsymbol{a}\rangle = \mathrm{f}(\bar{\boldsymbol{g}},\boldsymbol{a}), \qquad (6)$$

where $\bar{\boldsymbol{g}}$ is a vector representing the *average grader*,

$$\bar{\boldsymbol{g}} = \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{g}\in\mathcal{G}} \boldsymbol{g}.$$

In order to simultaneously consider individual and final grades, we needed to find the embedding matrices $\boldsymbol{W}$ (3) and $\boldsymbol{V}$ (4) that would give rise to a ranking as similar as possible to the ranking provided by graders, which we did by optimizing the function

$$\mathrm{f}(\bar{\boldsymbol{g}},\boldsymbol{a}) + \mathrm{f}(\boldsymbol{g},\boldsymbol{a}) = \langle \boldsymbol{W}\bar{\boldsymbol{g}}, \boldsymbol{V}\boldsymbol{a}\rangle + \langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a}\rangle = \\ \langle \boldsymbol{W}(\bar{\boldsymbol{g}}+\boldsymbol{g}), \boldsymbol{V}\boldsymbol{a}\rangle = \mathrm{f}(\bar{\boldsymbol{g}}+\boldsymbol{g},\boldsymbol{a}). \qquad (7)$$

This optimization maximized the similarity of rankings, measured by computing the proportion of pairs of answers whose relative order was the same. We thus used the area under the ROC curve (AUC), also known as the *concordance index* (C-index), pairwise ranking accuracy, or the *Kendall-τ* distance [5].

In symbols, the similarity of a grading function h with the ranking recorded in $\mathcal{D}$ is given by:

$$\mathrm{AUC}(h,\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{g},\boldsymbol{a}_b,\boldsymbol{a}_w)\in\mathcal{D}} \mathbb{I}_{h(\boldsymbol{g},\boldsymbol{a}_b)>h(\boldsymbol{g},\boldsymbol{a}_w)} + \frac{1}{2}\mathbb{I}_{h(\boldsymbol{g},\boldsymbol{a}_b)=h(\boldsymbol{g},\boldsymbol{a}_w)}. \qquad (8)$$

The symbol $\mathbb{I}_{(\bullet)}$ stands for the return of an indicator that returns 1 when a predicate $(\bullet)$ is true, and 0 otherwise.

Note that since this measure is not symmetric, when comparing two rankings we have to explicitly consider one as the *ground truth* and the other as the predicted ranking. In (8) we evaluate the quality of the ranking induced by h by considering that the preference judgements in D are drawn from the true ranking.

Tying up all the loose ends, the learning process aimed to optimize the embedding matrices in such a way that the individual and final grades were as coherent as possible with graders' rankings. Since the AUC (8) is not a convex function, we followed a maximum margin approach, defining:

$$\mathrm{err}(\boldsymbol{W},\boldsymbol{V}) = \sum_{(\boldsymbol{g},\boldsymbol{a}_b,\boldsymbol{a}_w)\in\mathcal{D}} \max(0, 1 - \mathrm{f}(\bar{\boldsymbol{g}}+\boldsymbol{g},\boldsymbol{a}_b) + \mathrm{f}(\bar{\boldsymbol{g}}+\boldsymbol{g},\boldsymbol{a}_w)). \qquad (9)$$

The goal was to ensure a difference of at least 1 for the sum of individual and final grades estimated for $\boldsymbol{a}_b$ and $\boldsymbol{a}_w$. To train the parameters that minimized (9) we used a stochastic gradient descent (SGD) algorithm, which iteratively updated, until a convergence criterion was reached, the model parameters as indicated in the following equation:

$$\Theta \leftarrow \Theta - \gamma \cdot \frac{\partial err}{\partial \Theta}, \tag{10}$$

where alternatively Θ is $\boldsymbol{W}$ and $\boldsymbol{V}$, and where γ is the *learning rate*. The partial derivative with respect to $\boldsymbol{W}$ when

$$(1 - \mathrm{f}(\bar{\boldsymbol{g}} + \boldsymbol{g}, \boldsymbol{a}_b) + \mathrm{f}(\bar{\boldsymbol{g}} + \boldsymbol{g}, \boldsymbol{a}_w) > 0)$$

(otherwise, 0), is given by:

$$\frac{\partial \mathrm{err}(\boldsymbol{W}, \boldsymbol{V})}{\partial \boldsymbol{W}} = \boldsymbol{V}\boldsymbol{a}_w(\bar{\boldsymbol{g}} + \boldsymbol{g})^T - \boldsymbol{V}\boldsymbol{a}_b(\bar{\boldsymbol{g}} + \boldsymbol{g})^T = \boldsymbol{V}(\boldsymbol{a}_w - \boldsymbol{a}_b)(\bar{\boldsymbol{g}} + \boldsymbol{g})^T. \tag{11}$$

The equation for the derivative with respect to $\boldsymbol{V}$ is analogous.

3.1. Grading graders

We define the *grade of a grader* $\boldsymbol{g}$ as the similarity between the his/her grades as predicted by the model, and the final grades, which are predicted by the mean grader. The ordering of the answers induced by $\boldsymbol{g}$ (5) depends only on the orientation of the associated vector in $\mathbb{R}^k$, that is, $\boldsymbol{W}\boldsymbol{g}$, whereas the final grades depend solely on $\boldsymbol{W}\bar{\boldsymbol{g}}$.

The grade of a grader (GG) is, thus, defined as the cosine of the angle between $\boldsymbol{W}\boldsymbol{g}$ and $\boldsymbol{W}\bar{\boldsymbol{g}}$ in the common Euclidean space,

$$\mathrm{GG}(\boldsymbol{g}) = \cos(\boldsymbol{W}\boldsymbol{g}, \boldsymbol{W}\bar{\boldsymbol{g}}) = \frac{\langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{W}\bar{\boldsymbol{g}} \rangle}{\|\boldsymbol{W}\boldsymbol{g}\| \cdot \|\boldsymbol{W}\bar{\boldsymbol{g}}\|}. \tag{12}$$

4. Experimental setting and results

Below we describe a pilot peer-assessment experiment in a real-world setting carried out at three Spanish higher-education institutions: University of A Coruña (UDC), Pablo de Olavide University in Sevilla (UPO) and the University of Oviedo (Uniovi) in Gijón. The data is the same as that used for previous research [7] for a non-content-based approach (that is, the words in the answers were not used in any way). The scenario of the experiment was the following. Undergraduate students of the Intelligent Systems (Computer Science) course were asked to write an essay answering some basic questions on informed and uninformed search methods. Using tools developed by Knoll et al. [33], the students were asked to use a search prototype to find the shortest paths in a small graph representing Vancouver neighborhoods. The students had to use three algorithms implemented in the search prototype to complete a table with details of the lengths of the best paths and the number of nodes expanded by each algorithm in each search. Since assessment of this question could be easily automated, it was discarded from the peer-assessment experiment. Students

Table 1: Main characteristics of the peer assessment experiment. The sparseness factor indicates the percentage of missing grades in the assessment matrix.

# of graders	160
# of answers	175
# of evaluations	1326
Sparseness (%)	95.26
# of evaluations per grader, mean	8.29 ± 1.45
# of evaluations per answer, mean	7.58 ± 2.02

next were asked to discuss the results obtained and their answers had to justify the results according to the optimality of the algorithms used.

The students had to anonymize their answers before submission to an event called JRLO2014 (Joint Research in Learning to Order 2014) created in EasyChair. After answers were collected by the EasyChair application, the students, acting as reviewers (graders), were given a small number of answers, together with a detailed rubric that explained how to assess the answers on a numeric scale from 0 (worst) to 10 (best). Graders were chosen at random and no student received their own answer to evaluate.

Answers were also evaluated by three instructors (one per university). Two instructors were designers of the experiment; the third instructor was provided with the rubric and with an explanation of the experimental process.

Below we statistically describe the dataset and then report results for a comparison of the method described in Section 3 with a baseline approach and with the evaluations of the instructors and intended to measure the impact of using words in the vectorial representation of the answers. The final subsection reports the scores obtained with the method described in Section 4.4.

4.1. Dataset

The peer assessment process provided us with grades as awarded by students of the three universities acting as graders. A total of 175 students submitted an answer each and a subset of 160 students participated as graders. Each student graded an average of 8.29 answers and each answer was graded an average of 7.58 times. The total number of grades collected was 1326 (Table 1). The sparseness of the assessment matrix is explained by the fact that only 4.74% of a total of $160 \times 175 = 28000$ possible grades (if every grader had evaluated all the answers) were obtained.

Table 2 shows the main statistical properties of grade distribution in the dataset. Note that the quality of the assessments would be quite poor if we were trying to use them in a cardinal sense. Figure 1 shows histograms with information about the grades awarded to the answers. Notice that the grade distribution mode is 5 points.

We built a set $\mathcal{D}$ of preference judgements (2) using the grades awarded in the peer assessment process to construct triples (as explained in Section 3). We

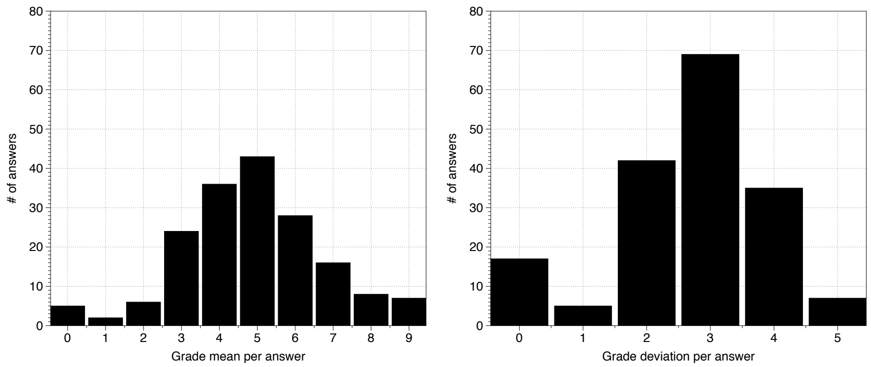


Figure 1: Grade mean and deviation per answer.

Table 2: Basic dataset statistics. The numeric grades must be in the range $[0, 10]$.

	Average
grade	4.84 ± 3.39
assessment range given per grader	7.31 ± 3.00
assessment range received per answer	6.67 ± 3.25

used simple binary coding for the graders in each triple (grader and answers) and represented the answers in two ways. The simplest one did not use words at all; rather, each answer was represented by the binary code for one identifier. However, we tried in this research to take advantage of the set of words present in each answer, so we also used a shallow natural language processing to include the answers' content as part of the input vector. We also borrowed techniques from the information retrieval field, such as the term-document matrix, $\boldsymbol{T}$, which represents the occurrence of terms in columns in a set of documents in rows. This matrix was built after parsing the answers of the students. No stemming or stop-word list was used.

We built the following extended dataset based on the preference judgements of the original dataset $\mathcal{D}$:

$$\mathcal{D}^{words} = \{[\boldsymbol{g}, \boldsymbol{a}_b \oplus \boldsymbol{T}_b, \boldsymbol{a}_w \oplus \boldsymbol{T}_w] : [\boldsymbol{g}, \boldsymbol{a}_b, \boldsymbol{a}_w] \in \mathcal{D}\},$$

where $\boldsymbol{T}_i$ refers to the i-th row of the term-document matrix $\boldsymbol{T}$.

4.2. Experiment settings: input data and algorithm parameters

From dataset $\mathcal{D}$ we obtained a model capable of ranking answers according to the partial orderings given by the graders. This optimization task was addressed using SGD (10) to find the parameters $\boldsymbol{W}$ and $\boldsymbol{V}$ that minimized (9).

Table 3: Comparative of the rankings produced by instructors' grades measured by the AUC (8); thus to compare two rankings, the higher, the more similar. In each row, one of the instructor grades are considered the *ground truth*, for this reason the AUC in the main diagonal of the table is 1.0.

	Instructor		
Ground truth	UDC	Uniovi	UPO
UDC	1.000	0.618	0.625
Uniovi	0.589	1.000	0.687
UPO	0.650	0.795	1.000

The SGD was applied using a learning rate defined in terms of the iteration i by

$$\gamma \leftarrow \frac{1}{(\gamma_s \cdot i) + 1}.$$

The parameters used were the results of a grid search in the following ranges of values:

$$k \in \{2, 10, 50, 100\}, \quad \gamma_s \in \{10^e : e = -7, \dots, 0\},$$

where k is the dimension of the common space, $\mathbb{R}^k$, where both graders and answers are mapped, and where γ_s regulates the reduction speed of the learning rate γ.

Using this model, a predicted ranking was obtained by applying the utility function of the *average grader*, as shown in (6), in order to obtain the final grade for each answer. This is, in fact, equivalent to filling the assessment matrix (1) by estimating the grade for each possible {grader, answer} pair and then computing a final grade as the mean of each column in $\boldsymbol{M}$.

Note that the output of the utility function could not be used directly as a grade since it was not bounded to any range of values. However, this output could eventually be transformed into valid grades by interpolating, for instance, between the grades provided by the instructors. We did not need to make this transformation, however, because in this study we are only interested in the ranking of answers.

In order to check the sensitivity of our approach with respect to the number of answers supplied to graders to evaluate, we built new datasets for each question by sampling the original data with different maximum numbers of answers per grader (5 to 11).

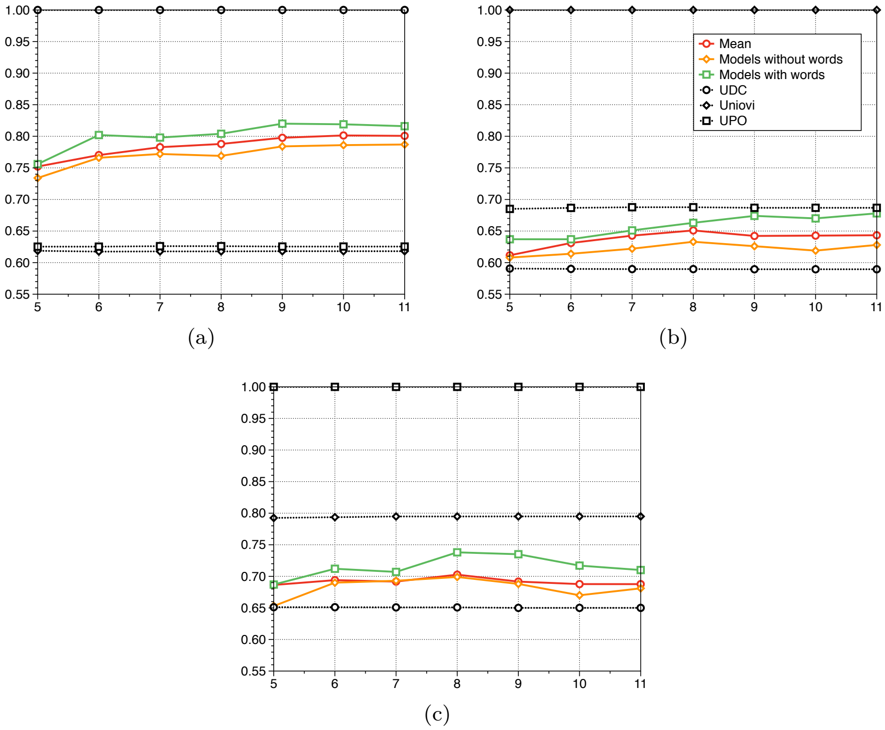
Finally, given that the SGD is a random algorithm, we repeated each experiment 10 times and averaged the scores (Tables 4 and 5).

4.3. Performance comparison

We compared the rankings of our method (both using and not using the words of the answers), of a baseline algorithm – obtained by averaging the grades awarded by each grader – and of the instructors. Note that the instructors were required to evaluate all the answers, not just those of their own students.

Table 4: Performance considering that the *ground truth* are the grades given by professional instructors of UDC, Uniovi and UPO. The scores are measured by the AUC (8); thus, higher values mean more similar rankings.

Ground truth	Maximum answers per grader	Mean	Models No words	Words
UDC	5	0.752	0.734	0.756
	6	0.770	0.766	0.802
	7	0.783	0.772	0.798
	8	0.788	0.769	0.804
	9	0.798	0.784	0.820
	10	0.801	0.786	0.819
	11	0.801	0.787	0.816
Uniovi	5	0.612	0.608	0.637
	6	0.631	0.614	0.637
	7	0.643	0.622	0.651
	8	0.651	0.633	0.663
	9	0.642	0.626	0.674
	10	0.643	0.619	0.670
	11	0.643	0.628	0.678
UPO	5	0.686	0.653	0.687
	6	0.694	0.690	0.712
	7	0.692	0.693	0.707
	8	0.703	0.699	0.738
	9	0.692	0.688	0.735
	10	0.688	0.670	0.717
	11	0.688	0.681	0.710



(a)

(b)

(c)

Figure 2: Graphical representation of the scores given in Table 4. The Y axis represents the AUC and the X axis represents the number of answers assessed per grader (not applicable to instructors scores). The ground truth is the ranking of UDC (a), Uniovi (b), and UPO (c).

Model performance was assessed in terms of the AUC (8). Since this is not a symmetric measure, we needed a ranking for comparison purposes and so we compared all rankings considering alternatively the ranking of each instructor as the ground truth.

We also compared the rankings of the instructors. Table 3 shows similarity scores for the instructors' rankings, taking the ranking given by the instructor in each row of the table as the *ground truth*. It can be seen that the UPO instructor achieved the highest scores and that the similarity between UPO and Uniovi instructors was generally greater than for any other instructor pairs and also that the UDC instructor grades were somewhat different.

Table 4 and Figure 2 show comparative results for the rankings by our algorithm and by the baseline algorithm, with the similarity of the rankings obtained by each approach measured against the instructor rankings. The scores all transmit a consistent outcome, namely, that the AUCs of models without words performed worst, those for the baseline algorithm performed moderately well, and those for models with words performed best. The instructors did not always perform better than the models. Therefore, even in the simplest case (without words), our approach was as good as or better than the instructors.

The scores for the baseline algorithm are slightly different from those reported by us elsewhere [7]. The reason is that previously we used all the answers available (175), whereas here we only used the answers of the 160 students who participated in the grading process. We think that the latter approach is fairer, given that the other methods only have access to those 160 answers.

Our scores are similar to other published results for peer-assessment experiments. Raman and Joachims [5], for instance, performed a similar experiment with two datasets, obtaining AUCs that ranged between 0.657 and 0.778 with respect to the grade rankings of teaching assistants. Note that the answers in that study were evaluated by 23.71 and 13.32 graders on average, compared with our average per answer of 7.58 graders.

Additionally, we observed – although not consistently – a slight increase in the AUC scores as the number of answers evaluated by each grader increased. This would indicate that the number of answers assessed by each grader does not need to be very high in order to obtain a reasonably accurate model.

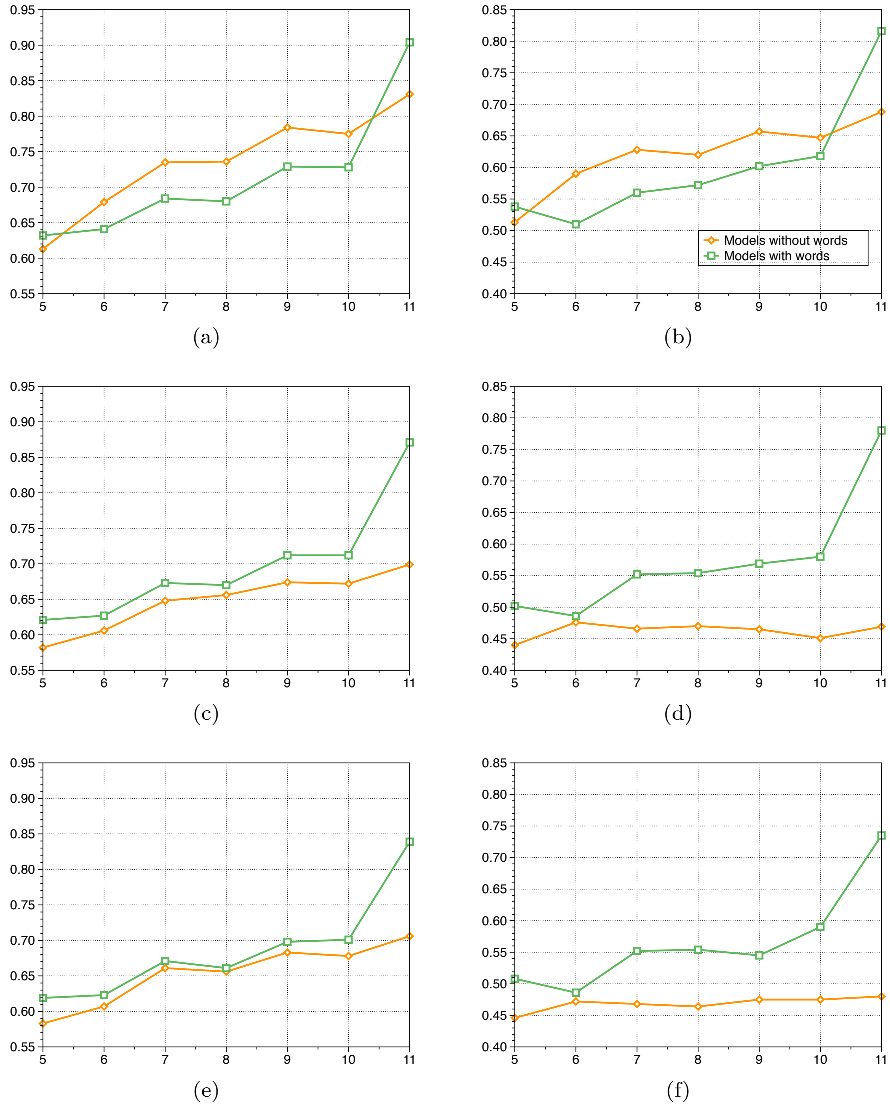
4.4. Grader grading

We evaluated the performance of the GG measure defined in (12) to grade graders. We again checked the similarity of two rankings: the ranking of graders as given by GG and the ranking of graders as obtained on comparing their assessments with those given by one of the instructors. Table 5 and Figure 3 show the results for these comparisons, reporting the AUC and the proportion of coincidences in the first tercile to highlight the coherence of the rankings.

Since the key issue was how to establish the *true* ranking of graders considering the grades awarded by an instructor, we considered the model trained with all available assessments (each grader evaluated a maximum of 11 answers) and compared the individual grader grades (5) with the instructor ranking. Thus,

Table 5: Comparison of the ranking yielded by GG (12) and the ranking resulting from a comparison of assessments by graders with assessments by one instructor. The comparisons were measured using the AUC for rankings and the proportion of common elements in the first tercile (T1).

Ground	Maximum answers	No Words		Words	
truth	per grader	AUC	T1	AUC	T1
UDC	5	0.613	0.513	0.632	0.538
	6	0.679	0.590	0.641	0.510
	7	0.735	0.628	0.684	0.560
	8	0.736	0.620	0.680	0.572
	9	0.784	0.657	0.729	0.602
	10	0.775	0.647	0.728	0.618
	11	0.831	0.688	0.904	0.816
Uniovi	5	0.582	0.440	0.621	0.502
	6	0.606	0.476	0.627	0.486
	7	0.648	0.466	0.673	0.552
	8	0.656	0.470	0.670	0.554
	9	0.674	0.465	0.712	0.569
	10	0.672	0.451	0.712	0.580
	11	0.699	0.469	0.871	0.780
UPO	5	0.583	0.446	0.619	0.508
	6	0.607	0.472	0.623	0.486
	7	0.661	0.468	0.671	0.552
	8	0.656	0.464	0.661	0.554
	9	0.683	0.475	0.698	0.545
	10	0.678	0.475	0.701	0.590
	11	0.706	0.480	0.839	0.735



(a) (b) (c) (d) (e) (f)

Figure 3: Graphical representation of the scores given in Table 5. The Y axis represents the AUC and the X axis represents the number of answers assessed per grader. The first and second columns depict the AUC and the proportion of common elements in the first tercile (T1), respectively. Rows represent the scores considering each instructor as the ground truth: UDC (a, b), Uniovi (c, d), and UPO (e, f), respectively.

for a given instructor, we had a ranking of graders that we could consider to be the ground truth.

A first interpretation of the results (Table 5) indicates that the actual number of answers evaluated by each grader was significant, as we could not expect to obtain a faithful measure of grader quality if the grader had only evaluated a few answers. We obtained AUCs of above 0.7 in most cases for the assumption that each grader evaluated at most 9 answers and the proportion of common graders in the first tercile was above 0.55. Figure 3 graphically represents these scores.

Observed in all cases was significant growth in both AUCs and in tercile proportions when all 11 grades awarded by the students were taken into account, attributable to the way we estimated the true ranking of graders given an instructor. Note that this increase was always lower when words were not used to represent the answers.

The use of words greatly increased scores in comparisons with UPO and Uniovi instructor scores but not when UDC instructor grades were considered as ground truth. The reason, already noted, is that the UDC instructor's assessments were significantly different from the other two instructors' assessments, probably due to a different interpretation of the rubric, which may not have been explained to the UDC instructor — on an assumption of tacit knowledge — as carefully as to the students.

5. Summary and conclusions

We described a factorization method to address the assessment of open-response answers in peer-grading contexts like those reflected in MOOCs. Our method uses a scalable SGD optimizer to train a scoring (utility) function to rank answers better than a baseline method consisting of averaging peer grades. To avoid the subjectivity of numeric scores, the learning algorithm starts from a collection of preference judgements and optimizes ranking by minimizing AUC error using a maximum margin approach.

The proposed method was tested in a real-world university setting using datasets collected in a peer-grading experiment. The impact of using additional information to describe the answers was evaluated by studying outcomes when shallow natural language processing of students' answers was included. The steps in our experiment can be summarized as follows:

1. We built a dataset from grades awarded by graders (1) following specifications for preference judgements (2). Both answers and graders were represented by feature vectors using binary coding for identification purposes.
2. We computed the term-document matrix from the open-ended answer texts and annotated the occurrence of terms in all the answers.
3. We enlarged the original dataset to include the term-document matrix representation of answers.

4. We optimized the error function (9) using an SGD (10) so as to obtain, as optimal parameters, the embedding matrices $\boldsymbol{W}$ (3) and $\boldsymbol{V}$ (4).

The results of a comparison of the rankings of answers by our method and by the three university instructors confirm that it is possible to obtain reliable rankings using an ordinal approach to peer assessment. The main contribution of our research, however, is that it demonstrates that using content-based elements notably enhances peer grading performance. Including a simple vectorial representation of texts provided by students as answers improves the quality of the assessments used for model training.

We also checked the validity of a method to grade graders by comparing it with rankings induced by the coherence of grader assessments with instructor assessments. Again, the use of answer content generally improved our results.

As future research we plan to consider the usefulness of this approach in reducing the burden on students of peer-grading all answers in a course. If only a proportion of students graded a proportion of answers, those answers could be labeled with the grade awarded by the model and a simple text categorization approach could be used to grade the remaining answers. This would only be possible, however, if a content-based method like that described above were used.

Acknowledgment

This research was supported in part by the Spanish Ministerio de Economía y Competitividad (grants TIN2011-23558, TIN2012-37954, TIN2014-55894-C2-2-R, TIN2015-65069-C2-1-R, TIN2015-65069-C2-2-R), the Junta de Andalucía (grant P12-TIC-1728) and the Xunta de Galicia (grant GRC2014/035), all, in turn, partially funded by FEDER. We would also like to thank the students from the University of A Coruña, Pablo de Olavide University and University of Oviedo who participated in this research.

References

[1] P. M. Sadler, E. Good, The impact of self-and peer-grading on student learning, Educational Assessment 11 (2006) 1–31.

[2] C. Kulkarni, K. Pang-Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, S. R. Klemmer, Peer and Self Assessment in Massive Online Classes, Technical Report, Stanford University, 2013.

[3] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, D. Koller, Tuned models of peer assessment in MOOCs, in: Proceedings of the 6^{th} International Conference on Educational Data Mining (EDM'13), International Educational Data Mining Society, 2013, pp. 153–160.

[4] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, K. Ramchandran, A case for ordinal peer-evaluation in MOOCs, in: NIPS Workshop on Data Driven Education, 2013.

[5] K. Raman, T. Joachims, Methods for ordinal peer grading, in: ACM Conference on Knowledge Discovery and Data Mining (KDD), 2014.

[6] J. Díez, O. Luaces, A. Alonso-Betanzos, A. Troncoso, A. Bahamonde, Peer Assessment in MOOCs Using Preference Learning via Matrix Factorization, in: NIPS Workshop on Data Driven Education, 2013.

[7] O. Luaces, J. Díez, A. Alonso-Betanzos, A. Troncoso, A. Bahamonde, A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments, Knowledge-Based Systems 85 (2015) 322 – 328.

[8] K. Raman, T. Joachims, Bayesian ordinal peer grading, in: Proceedings of the Second (2015) ACM Conference on Learning Scale, ACM, New York, NY, USA, 2015, pp. 149–156.

[9] T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2002.

[10] W. Barnett, The modern theory of consumer behavior: Ordinal or cardinal?, The Quarterly Journal of Austrian Economics 6 (2003) 41–65.

[11] A. Bahamonde, G. F. Bayón, J. Díez, J. R. Quevedo, O. Luaces, J. J. del Coz, J. Alonso, F. Goyache, Feature subset selection for learning preferences: A case study, in: Proceedings of the International Conference on Machine Learning (ICML '04), 2004, pp. 49–56.

[12] J. J. del Coz, G. F. Bayón, J. Díez, O. Luaces, A. Bahamonde, C. Sañudo, Trait selection for assessing beef meat quality using non-linear SVM, in: Advances in Neural Information Processing Systems 17 (NIPS '04), 2005, pp. 321–328.

[13] J. A. Krosnick, Survey research, Annual Review of Psychology 50 (1999) 537–567.

[14] B. Carterette, P. N. Bennett, D. M. Chickering, S. T. Dumais, Here or There. Preference Judgments for Relevance, in: Advances in Information Retrieval, Springer, 2008, pp. 16–27.

[15] R. Herbrich, T. Graepel, K. Obermayer, Support vector learning for ordinal regression, in: Proceedings of the Ninth International Conference on Artificial Neural Networks, Edinburgh, UK, 1999, pp. 97–102.

[16] O. Luaces, J. Díez, A. Alonso, A. Troncoso, A. Bahamonde, Including content-based methods in peer-assessment of open-response questions, in: Proceedings of 2015 IEEE 15th International Conference on Data Mining Workshop Data Mining for Educational Assessment and Feedback (ASSESS 2015), 2015, pp. 273–279.

[17] J. Díez, O. Luaces, A. Alonso-Betanzos, A. Troncoso, A. Bahamonde, Calificación de calificadores en la evaluación por pares de exámenes de respuesta abierta, in: Actas de la XVI Conferencia de la Asociación Española para la Inteligencia Artificial, VII Simposio de Teoría y Aplicaciones de Minería de Datos, Asociación Española para la Inteligencia Artificial, 2015, pp. 717–726.

[18] V. Aggarwal, S. Srikant, V. Shashidhar, Principles for using Machine Learning in the Assessment of Open Response Items: Programming Assessment as a Case Study, in: NIPS Workshop on Data Driven Education, 2013.

[19] N. B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, M. J. Wainwright, Some scaling laws for MOOC assessments, in: KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014), 2014.

[20] D. Pérez-Marín, I. Pascual-Nieto, P. Rodríguez, Computer-assisted assessment of free-text answers, The Knowledge Engineering Review 24 (2009) 353 – 374.

[21] F. Rodrigues, P. Oliveira, A system for formative assessment and monitoring of students' progress, Computers & Education 76 (2014) 30 – 41.

[22] G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing, Communications of the ACM 18 (1975) 613–620.

[23] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[24] F. Noorbehbahani, A. Kardan, The automatic assessment of free text answers using a modified BLEU algorithm, Computers & Education 56 (2011) 337 – 345.

[25] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American society for information science 41 (1990) 391–407.

[26] P. Thomas, D. Haley, A. deRoeck, M. Petre, E-assessment using latent semantic analysis in the computer science domain: A pilot study, in: COLING 2004 eLearning for Computational Linguistics and Computational Linguistics for eLearning, Geneva, Switzerland, 2004, pp. 38–44.

[27] R. Blake, O. Gutierrez, A semantic analysis approach for assessing professionalism using free-form text entered online, Computers in Human Behavior 27 (2011) 2249 – 2262.

[28] D. Pérez, A. M. Gliozzo, C. Strapparava, E. Alfonseca, P. Rodríguez, B. Magnini, Automatic assessment of students' free-text answers underpinned by the combination of a bleu-inspired algorithm and latent semantic analysis., in: FLAIRS Conference, 2005, pp. 358–363.

[29] C. Gütl, Moving towards a fully automatic knowledge assessment tool, International Journal of Emerging Technologies in Learning 3 (2008).

[30] H.-C. Wang, C.-Y. Chang, T.-Y. Li, Assessing creative problem-solving with automated text grading, Computers & Education 51 (2008) 1450 – 1466.

[31] A. Spalvieri, S. Mandelli, M. Magarini, G. Bianchi, Weighting peer reviewers, in: Privacy, Security and Trust (PST), 2014 Twelfth Annual International Conference on, IEEE, 2014, pp. 414–419.

[32] D. Kurokawa, O. Lev, J. Morgenstern, A. Procaccia, Impartial peer review, in: IJCAI, 2015.

[33] B. Knoll, Kisyński, G. Carenini, C. Conati, A. Mackworth, D. Poole, AIspace: Interactive tools for learning artificial intelligence, in: Proceedings of the AAAI 2008 AI Education Workshop, Chicago, IL, 2008.