Including Content-Based Methods in Peer-Assessment of Open-Response Questions

Oscar Luaces, Jorge Díez
Artificial Intelligence Center,
University of Oviedo
33204 Gijón, Spain
oluaces@uniovi.es

Amparo Alonso
Dept. of Computer Science,
University of A Coruña
15071 A Coruña, Spain
ciamparo@udc.es

Alicia Troncoso
Dept. of Computer Science,
Pablo de Olavide University,
41013 Sevilla, Spain
atrolor@upo.es

Antonio Bahamonde
Artificial Intelligence Center,
University of Oviedo
33204 Gijón, Spain
abahamonde@uniovi.es

***Abstract*—Massive Open Online Courses (MOOCs) are attracting the attention of a huge number of students all around the world. These courses include different types of assignments in order to evaluate the student's knowledge. However, these assignments are designed to allow a straightforward automatic evaluation. But, in this way it is not possible to evaluate some skills that would require answering open-response questions. Peer-assessment (the students are asked to assess other assignments), is an effective method to overcome the impossibility of having staff graders for this task. Additionally, students gain a deeper knowledge about the subject of the assignments that they have to read critically. However, the grades given by student-graders must be filtered to avoid bias due to a lack of experience in assessment tasks. There are a number of approaches to do this. In this paper we present a factorization approach that in addition to the grades given by graders is able to incorporate a representation of the contents of the responses given by students using a Vector Space Model of the assignments. So we fill the gap between peer-assessment and content-based methods that use a shallow linguistic processing. The paper includes a report of the results obtained using this approach in a real world dataset collected in 3 universities of Spain, A Coruña, Pablo de Olavide at Sevilla, and Oviedo at Gijón. The scores obtained by the method presented here were compared with those provided by the staff of these universities. We report a considerable improvement whenever we use the content-based approach. In any case, we conclude that there is no evidence that staff grading would have led to more accurate grading outcomes than the assessment produced by our models.**

I. Introduction

Massive Open Online Courses (MOOCs) have attracted thousands of students from many parts of the world. These courses promise new educational possibilities and have focused the attention of many researches to improve the education experience of students. Assessment is, in general, an important part of the teaching process, and it has to be addressed in order to provide a feedback to students and to guarantee the quality of the titles given to graduates. In this paper we tackle the challenge of evaluating open-response questions. Notice that when there is a very large number of assignments, it is imposible to be assessed by instructors or teaching assistants (TA).

We adopt the approach called peer-assessment [1], [2], [3], [4], [5], [6], [7], [8] as the basic strategy. The students are asked to evaluate a small set of anonymized assignments submitted by other students. Additionally, these student-graders receive a set of detailed rules (called *rubric*) in order to uniform the assessment. However, students typically have no experience in this task and then effective peer grading must deal with the effects of inconsistent subjective evaluation.

It is important to remark here that peer-assessment has a pedagogic value *per se*. When students are asked to assess a number of assignments, they must read and compare the answers of other students and this process produces a deep understanding of the contents of the course.

There are two main streams in peer-assessment: *cardinal* and *ordinal*. In the first case, grades are numbers or categorical labels with a straightforward numerical semantics. The assessment returned by student-graders are cardinal values. The semantic of these grades is clear for students since they have been receiving these feedback for years. In peer-assessment, the final grade given to an assignment is usually determined by an aggregation function, typically the average or median [2] given by graders.

The cardinal approach has some important flaws. The assessments are usually affected by some graders' bias that would deviate them with respect to the *ideal ground truth*. The presumed universality of the semantics of cardinal grades is not so general. Some students tend to give high grades, while others (probably with different academic backgrounds) are less generous with their assessments. If we had a large number of grades for each assignment, then, the *correct* grade could be approximated by averaging all available grades. In this case it has been reported [3] that averages are more consistently accurate with respect to the rubric than the staff grades. However, it is not always possible to reach this ideal situation, as students can not be charged with the job of grading a large number of assignments. Unfortunately, we can only obtain a few assessments for each assignment.

In addition to the misunderstanding of the semantic of grades, there is a second shortcoming that has to be addressed in cardinal approaches: the *batch effect*. It has been observed [9], [10] that an item tends to receive a higher grade when it is evaluated in a batch of worse items than when it is evaluated in a group of better items. To overcome the disadvantages of cardinal assessment, we may use the ordinal approach. Instead of asking graders to assign a cardinal value, they are asked to provide a ranking of the assignments that they have to assess. This is an easier task for inexpert graders, and the reliability of the rankings is considerably higher than that of the cardinal assessments [11], [9], [1], [12]. In a context of information retrieval, in [13] the authors propose a preference approach to learn the relevance of documents. An interesting discussion about cardinal and ordinal from a psychological point of view is presented in [14].

On the other hand, there are content-based assessment methods. In the next section we review these approaches. All of them use Information Retrieval techniques including shallow linguistic processing. Content-based methods require some help from the instructor. Sometimes, several ideal answers (references) are compared with the answers given by the students. Other times, a subset of assignments labeled with the *correct* grades is extended to the whole set of assignments using a Machine Learning algorithm.

No known peer-assessment method, ordinal or cardinal, takes into account the contents of the open answers given by the students. Nevertheless, somehow peer-assessment methods work like collaborative filters that aim to *recommend* a grade to each assignment. Thus the techniques used by recommender systems could be adapted to the assessment task. In this paper we present an approach that tries to combine the strengths of ordinal collaborative filters and content-based recommenders. We use a factorization method to learn a utility function that estimates the consensus ordering of the assignments. This approach was used in our previous work [6], and it is inspired by the framework presented in [15], [10] to *learn preferences*.

The assignments can be represented by feature vectors. The use of features has been acknowledged to be crucial for the success in some cases; see the work of [16] and [17]. If no other information is available, the features just capture a binary identification of assignments and graders. In such case we have only a pure collaborative approach. But the factorization method presented here allows representations including any information about the assignments. Let us underscore that in the method presented here it is not needed (as in other approaches) any self-grading of the assignments nor any previous gradings by instructors.

After the formal presentation of the assessment method, the paper is closed with the report of the results obtained with two real world datasets obtained from a common assignment for Computer Science students of 3 universities (in this anonymized version we call them): of Spain: A Coruña, Pablo de Olavide at Sevilla, and Oviedo at Gijón. We found that our method achieves similar or better scores than staff instructors when we measure the discrepancies with other instructors' grades. We tested both collaborative filtering and content-based representations. We checked that content-based version achieves considerably better results when the scores are compared with the assessments given by staff instructors.

II. Related Work

There are many related works in this area; some of them have just been mentioned in the Introduction. As it happens in Recommender Systems, automatic assessments can be divided in two groups, those that use the contents of the answers provided by students and those that work like a collaborative filtering. The approach presented in this paper can be seen as a collaborative filtering extended to capture the contents of the assignments.

In the content-free stream, the most similar work is [5]. The authors report a case study with real data from a Cornell University course. The assignments are 42 posters and 44 reports done by groups of students. Each poster received an average of 23.71 grades, while reports received 13.32 grades. The authors propose to use the ordinal approach casting the learning problem as a *rank aggregation* learning task. The paper compares the performance of several probabilistic aggregation algorithms and acknowledges that *simply averaging the cardinal scores of the peer graders performs surprisingly well*. Probably the reason is that each assignment receives a high number of grades in coherence with results reported for instance by [3]. The paper compares the accuracy of the models learned with the rankings achieved by a set of teaching assistants (TA). The conclusion is that *there is no evidence that TA grading would have led to more accurate grading outcomes than peer grading*.

Another probabilistic learning algorithm has been proposed for peer grade estimation by [3]. The paper presents also a case study with 63199 peer grades of a Coursera course about Human Computer Interaction (HCI). Their method requires self-grading of the students and the evaluation of some assignments that were previously graded by the instructor in order to estimate grader reliability. Our method is not constrained by these requirements.

Both papers, [3], [5], emphasize the relevance of assessing the accuracy of graders. In fact, it is crucial to incentive students to make a good evaluation if we want to obtain reliable data. A way to do this is to include the assessments carried out by the students as part of their final grade. On the other hand, we think that the evaluation process itself may be an additional way for students to get insight into the field covered in the assignment.

The authors of [4] use also a dataset of HCI on Coursera; in this case from the third offering of the course. The used dataset has assignments submitted by 1879 students, and 7242 numerical grades were collected by a peer grading experiment. The authors acknowledge as a desideratum to seek for a *trade-off between the precision of cardinal scores and the robustness of ordinal evaluations for peer grading*. The computational method proposed in the paper is an ordinal approach that searches for the solution of a non-convex optimization problem that uses a logistic sigmoid. The experiments reported achieve a performance similar to the performance of a method that simply computes the median of the grades given to each assignment.

In [17], the authors present formal proofs about the errors in peer-gradings when the grade is estimated averaging the grades given by graders. There is a constant proportion of assignments erroneously graded. The amount of assignments may become too high in MOOCs, and therefore the procedure is unacceptable. So, the proposal of the paper is to use methods that include some kind of *dimensionality reduction*; in particular, the authors discuss clustering and featuring. Although the proposals are very abstract, it is interesting to underscore that the factorization method proposed in our paper is a suitable framework to implement both approaches.

On the arena of content-based systems, the general idea is to use a combination of a shallow Natural Language Processing and Machine Learning. Somehow, the methods are borrowed from Information Retrieval. Roughly speaking, we may distinguish between matching and categorization methods.

Matching methods compare the students' answer against some reference (ideal answer) or template; [18] made a detailed survey of published algorithms using this paradigm.

To match the contents of the students' answers with the references, [19] compute a cosine similarity after a preprocess. Both, references and students' answers, are represented using the Vector Space Model (VSM), where each word is the index of a vector with values recording the presence or the frequency of the word in the document [20]. This values may be weighted using different strategies.

Some authors have used matching methods that exploit the coincidence of groups of words, in order to take into account the syntactic structure of the documents without penalizing the process with a deep analysis. Here, a key tool is the metric called BLEU [21]. This is a metric of document similarity devised to assess the quality of translations. Given a set of reference translations, BLEU computes a scoring for a candidate translation based on the co-occurrence of n-grams in any of the references and the candidate. A modification of BLEU is used by [22] to build an automatic assessment of open-ended answers.

The major disadvantage of these methods is that they do not consider synonyms. However, certain semantic analysis is necessary to make a fair comparison of students' answers and the references; we can not expect students to repeat exactly the same words used by the reference answers given by the instructors. To overcome this problem, one standard option is to use *Latent Semantic Analysis* (LSA) [23]. This method projects the matrix of VSM representations of all answers (usually called *term-document* matrix) into a smaller dimensional space using the Singular Value Decomposition (SVD) of the matrix. This method used in Information Retrieval is robust and captures the implicit semantic in the set of available documents.

A pilot study based on LSA was carried out by [24] to evaluate the answers from six students to three questions in the Computer Science domain; the paper reports a high precision despite the small size for the dataset. Again, LSA was applied by [25] for assessing the *professionalism*, depending on five attitudinal categories of free-form text responses from participants in a professional development program. In this work, a previous preprocessing step based on standardization, stop word removal and Porter stemming was applied to obtain the term-document matrix. On the other hand, [26] propose a combination between BLEU and LSA to assess open-ended answers.

We would like to emphasize that the factorization approach presented in this paper is a generalization of the matrix decomposition provided by SVD. In our approach the decomposition is obtained aiming at the optimization of a loss function, in order to improve the predicted outcome of the model. In the next section we give the details of this formulation.

Another approach that uses some kind of semantic analysis explicitly is presented by [19]. In addition to other options, the system presented in their paper makes an automatic assessment using an extension of the cosine similarity. It takes into account that two words are considered similar if they are related in the WordNet semantic network.

There is another group of approaches that use the contents of the assignments. This is an adaptation of text categorization. The instructor grades a reduced set of assignments and the completion of the job is done by an ordinal classifier learned from the supervised dataset of assignments and grades. In this context, the assignments are represented by a feature vector. In [27] the *CarmelTC* algorithm, that uses a Naïve Bayes classifier, is proposed. On the other hand, [28] presented a Support Vector Machine (SVM) to evaluate creative problem-solving from open-ended responses. A comparison between the results obtained by SVM, LSA and a standard regression method showed that SVM provided the highest correlation with the scoring of the instructor.

III. Learning Method

Let $\mathcal{G}$ be a set of *graders* and let $\mathcal{A}$ be a set of *assignments*. Each grader g has received a subset $\mathcal{A}_g \subset \mathcal{A}$ of assignments to evaluate. The initial data to infer a grading function is the *assessment matrix*, $\boldsymbol{M}$, which contains the scores given by the 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 assignments will be represented by vectors of features; we will use the same symbols to name their vectorial representation or the grader or assignment. In the simplest case, a grader (respectively, an assessment) can be identified by a vector of binary values with all zeros but one 1 in the component indexed by itself in $\mathcal{G}$ (respectively $\mathcal{A}$). This simple representation can be enriched with features describing additional aspects of the graders/assessments. In Section IV we show the positive effect of this enriched representation in the case of assignments.

In general, matrix $\boldsymbol{M}$ is going to be very sparse. Only a few assignments are graded by each student. The goal of any peer-assignment method is to obtain an absolute ranking of assignments from the scores in $\boldsymbol{M}$. A straightforward way to do this is to rank assignments according to the average score for each assignment. This approach is accurate when each assignment has been assessed by a large number of graders, as we mentioned in the Introduction. However, this is not the case in a peer assessment context.

In our approach we are going to learn a scoring function able to fill the matrix $\boldsymbol{M}$ and then use the average scores of all graders on all assignments to obtain the final ranking. This scoring function is induced based on preference learning to avoid the subjectivity of graders. We will focus on the relative ordering of assignments for each grader, and not in the score values. Thus, we build a set of *preference judgments*, $\mathcal{D}$, given by triples of a grader $\boldsymbol{g}$ and a couple of assignments $(\boldsymbol{a}_b, \boldsymbol{a}_w)$ in $\mathcal{A}_g$ such that

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

Notice that assignments with the same score will not provide any relative order, so ties are discarded when generating the dataset of preference judgments.

The strategy that we propose to obtain the ranking starts with an double *embedding*: mapping both assignments and graders into a common Euclidean space $\mathbb{R}^k$:

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

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

The representation of assignments ($rep(\mathcal{A})$) may have a higher dimension than the number of assignments when we use the vectorial representation of the answers in addition to the identifier of the answer itself.

From dataset $\mathcal{D}$ and with the embeddings, we will define the *individual assessment* as a function from graders and assignments 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 assignment $\boldsymbol{a}$, it will be used to complete the assessment matrix. Then, we can compute the *final grade* for each assignment as the average of all its grades.

$$\begin{aligned} \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \mathrm{f}(\boldsymbol{g}, \boldsymbol{a}) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a} \rangle = \\ \Big\langle \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \boldsymbol{W}\boldsymbol{g}\boldsymbol{V}\boldsymbol{a} \Big\rangle = \langle \boldsymbol{W}\bar{\boldsymbol{g}}, \boldsymbol{V}\boldsymbol{a} \rangle = \mathrm{f}(\bar{\boldsymbol{g}}, \boldsymbol{a}), \end{aligned} \tag{6}$$

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 consider at the same time individual and final grades, we are trying to find the embedding matrices $\boldsymbol{W}$ and $\boldsymbol{V}$ (Eq. 3,4) that give rise to most similar ranking with those provided by graders. In a sense that we are going to explain next, we optimize the function

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

First, let us fix that the comparison of two rankings is going to be computed using the proportion of pairs of assignments which relative order is the same. That is to say, we use the area under the ROC curve (AUC). It is also known as the *concordance index* (C-index), or the pairwise ranking accuracy. This measure is called *Kendall-τ* in [5].

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

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

This measure is not symmetric, so when comparing two rankings we have to explicitly consider one of them as the *ground truth* and the other as the predicted ranking. In (Eq. 8) we evaluate the quality of the ranking induced by h considering that the preference judgments in $\mathcal{D}$ represent the true ranking.

Tying up all the loose ends, the aim of the learning process devised to make the assessment is to optimize the embedding matrices in such a way that the individual plus the final grades be as coherent with graders' orderings as possible. Since the AUC (Eq. 8) is not a convex function, we will follow a maximum margin approach. Then, we define

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

The idea is to ensure that the difference of sum of individual and final grades estimated for $\boldsymbol{a}_b$ and $\boldsymbol{a}_w$ is at least 1. To learn the parameters that minimize the previous equation we use a Stochastic Gradient Descent (SGD in the following) algorithm. The SGD approach iteratively updates the parameters of the model as indicated in the next equation, until a convergence criterion is reached;

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

where alternatively Θ is $\boldsymbol{W}$ and $\boldsymbol{V}$, and 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 is 0), is given by:

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

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

IV. Experimental settings and results

In this section we report a pilot experiment of peer assessment in a real-world context. It was carried out with the collaboration of three higher education institutions of Spain: University of A Coruña (UDC), University Pablo de Olavide at Sevilla (UPO), and University of Oviedo (Uniovi) at Gijón. The acronyms between parentheses, which appear in the figures and tables of results, come from the Spanish names of our universities.

The scenario of the experiment was the following:

- All the undergraduate students of the course *Intelligent Systems* (Computer Science) had to write an essay answering some basic questions about *informed and uninformed searching methods*.
    - Using the handbook [29], the students were asked to use a searching prototype in order to find the shortest paths in a small graph representing the neighborhoods of Vancouver. The students had to use 3 given algorithms, already implemented in the searching prototype, to fill a table with the lengths of the best paths and the number of nodes expanded in each search by each algorithm. The assessment of this question can be easily automated, so we discarded this question from the peer assessment experiment.

Table I. SOME CHARACTERISTICS OF THE DATA GATHERED IN OUR PEER ASSESSMENT EXPERIMENT. THE SPARSENESS FACTOR INDICATES THE PERCENTAGE OF MISSING GRADES IN THE ASSESSMENT MATRIX

# of graders	160
# of assignment	175
# of evaluations	1326
sparseness (%)	95.26
avg. evaluations per grader	8.29 ± 1.46
avg. evaluations per assignment	7.58 ± 2.01

  ○ Then, the students had to discuss the results obtained in the previous question. Their answers had to justify the results according to the optimality of the algorithms used.

- The students had to anonymize their assignments previously to the submission to an event registered in EasyChair with the name *JRLO2014 (Joint Research in Learning to Order 2014).*
- Once the assignments were collected by the EasyChair application, the students, acting as reviewers (graders), were given a few assignments together with a detailed rubric spelling out how to assess them. Each question had to be graded in a numeric scale of integers from 0 (worst) to 10 (best). The graders were chosen at random, avoiding that any student received his/her own exercise to evaluate.
- Finally, the students received the feedback from the anonymous reviewers of their assignments. All the assignments were also evaluated by the three instructors implicated in the experiment, but their scores where used only for comparison purposes, and not for the learning task.

In the rest of this section we first present the datasets used in the experiments, and then we show the results obtained in a comparison of our method with a baseline approach, and with the evaluations of the instructors of the courses.

A. Dataset

The peer assessment process provided us with a record of the grades given by our students for the discussion question introduced above. From the students of the 3 universities that submitted 175 assignments, a subset of 160 participated in the experiment as graders too. Each student received an average of 8.29 assignments to evaluate, while each assignment received in average 7.58 grades. The total number of grades collected was 1326 (see Table I). Notice the sparseness of the assessment matrix, since we have only 4.74% from a total of $160 \times 175 = 28000$ possible assessments that would be obtained if every grader would have evaluated all the assignments.

Table II shows some statistical properties of the grade distribution in the dataset. Let us remark that the quality of the assessments is quite bad if we were trying to use them in a cardinal sense. Figure 1 depicts the histograms with information about the grades received by the assignments. Notice that the mode of the distribution of grades is 5 points.

We built a set $\mathcal{D}$ of preference judgments (Eq. 2) using the grades given by the peer assessment process to construct the triples as explained in Section III. The elements of each triple (grader and assignments) were represented by vectors that identify them with a simple binary codification.

Table II. BASIC STATISTICS OF THE DATASETS. 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 assignment	6.67 ± 3.25

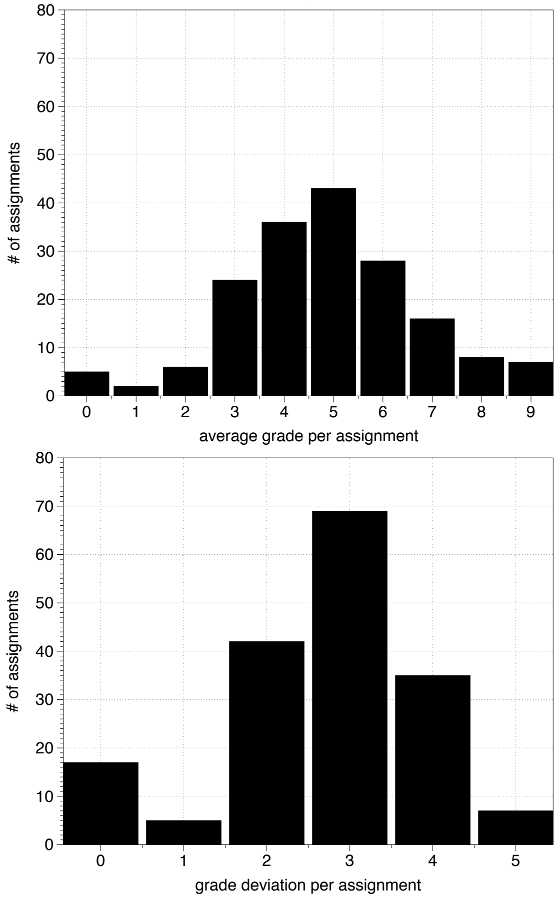

Figure 1. Average grade and deviation per assignment

B. Settings for the experiments: input data and algorithm parameters

Once we have the dataset $\mathcal{D}$, we obtained a model able to rank the assignments according to the partial orders given by the graders. This is an optimization task that was addressed using SGD (Eq. 10) to find the parameters $\boldsymbol{W}$ and $\boldsymbol{V}$ that minimize (Eq. 9).

The SGD was applied using in all cases 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:

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

Table III. Each column compares the ranking of each method with the ranking of one of the staff instructors. The scores shown are AUCs, thus the higher, the better. Of course, the instructor considered as the ground truth in each column reaches the best possible score, 1

Ground truth	UPO	Uniovi	UDC
model no words	0.679	0.627	0.785
model with words	0.718	0.671	0.830
mean	0.688	0.643	0.801
UPO	1	0.687	0.625
Uniovi	0.795	1	0.618
UDC	0.650	0.589	1

where k is the dimension of the common space where both graders and assignments are mapped, and γ_s regulates the reduction speed of the learning rate γ.

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

Notice that the output of the utility function is not bounded to any range of values, so it cannot be used directly as a grade. However, this output could eventually be transformed into valid grades using, for instance, some grades provided by professional instructors and interpolating. We do not need to make this transformation, because in this study we are only interested in the ranking of assignments.

C. Performance of the proposed method

We conducted a comparison among the ranking predicted by the method presented here, a baseline algorithm, and the rankings given by the instructors of the universities involved. The baseline ranking was obtained by averaging the grades given by each grader. The instructors' rankings required each instructor to evaluate all the assignments, not only those of their own students.

The performance of the models was assessed in terms of AUC (Eq.8). Let us recall that this is not a symmetric measure, so we need to fix a ranking to compare with. Thus we compared all rankings considering alternatively the ranking of each instructor as the *ground truth*.

The results are shown in Table III. Each column reports the AUCs that compare the ranking of each method with the ranking of one of the staff instructors considered as the ground truth in that column. Without considering the contents of the assignment in any sense, our method is better than the mean (baseline method). Moreover, the AUC obtained is better than the AUC obtained by at least one of the other two instructors. Therefore, even in the simplest case, the method presented in this paper is as good as (or better than) the profesional instructors.

These scores are similar to other published results on peer assessment experiments. For example, [5] carried out a similar experiment with two datasets, obtaining AUCs that ranged between 0.657 and 0.778 respect to the rankings given by TA (Teaching Assistant) grades. Worth of mention is that the assignments were evaluated on average by 23.71 and 13.32 graders, respectively, in their two datasets, while we have an average of 7.58 grades per assignment.

Next, we analyzed the impact of considering content-based information is the process. As was mentioned above, the simple binary codification used in the experiments described previously to plainly identify assignments can be extended to include additional features of the assignments. A rational extension for the representation of assignments would be to include somehow the answer given by the students. We used a shallow natural language processing to include the content of an answer as part of the input vector. We borrow some 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). It was built after parsing the answers written by the students. No stemming or stop-words list was used.

Then we built the following extended dataset based on the preference judgments in 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 $\boldsymbol{T}$.

The second row of Table III shows the results obtained with the extended representation. In all case the AUC is increased dramatically. Thus, we can conclude that the method proposed here can take advantage of the information about the contents of the answers.

V. Summary and conclusions

We have presented a factorization method to address the assessment of open-response assignments in peer-grading contexts, such as those used in MOOCs. Our method uses a scalable SGD optimizer that learns a scoring (utility) function able to rank the assignments better than a baseline method consisting of averaging the peer grades. The learning algorithm starts from a collection of preference judgments to avoid the subjectivity of the numeric scores, and optimizes the ranking by minimizing the AUC error using a maximum margin approach.

The proposed method was tested on a real-world setting, using datasets collected in a peer-grading experience carried out by three universities. With the data gathered, we analyzed the impact of using additional information to describe the assignments. More precisely, we studied the outcome produced when including a shallow natural language processing of the students' answers. In summary, the steps given in our experiment were:

1) We built a dataset from the scores given by the graders (Eq. 1) following the specifications of $\mathcal{D}$ as indicated in (Eq. 2). Both, assignments and graders, were represented by feature vectors using a binary codification of their identity.
2) We computed the term-document matrix from the text of the open-ended answers. We annotate the occurrence of the terms in all the assignments.

3) The original dataset were augmented to include the representation of assignments as given by the term-document matrix.

4) We used an SGD (Eq. 10) to optimize the error function (Eq. 9), thus obtaining the optimal parameters, the embedding matrices $\boldsymbol{W}$ (Eq. 3) and $\boldsymbol{V}$ (Eq. 4).

We showed the results of a comparison of the ranking of assignments produced by our methods and the rankings of the 3 staff instructors of the universities involved in the experiment.

These results, on the one hand, confirm that it is possible to produce reliable rankings in a peer assessment method using an ordinal approach. However, the contribution of the paper is that the use of content-based elements improves dramatically the performance of the method. Including a simple vectorial representation of the documents provided by students as answers to the assignments is very helpful in order to improve the quality the assessments. As a future work we are considering to analyze the impact of using other characteristics of the assignments beyond the text of the answers.

Acknowledgment

The research reported here is supported in part under grants TIN2011- 23558, TIN2012-37954 and TIN2011-28956-C02 from the MINECO (Ministerio de Economía y Competitividad, Spain), all partially supported with FEDER funds. We would also like to thank our students for grading the assignments of other colleagues in the Universities of A Coruña, Pablo de Olavide, and Oviedo.

References

[1] P. M. Sadler and E. Good, "The impact of self-and peer-grading on student learning," *Educational assessment*, vol. 11, no. 1, pp. 1–31, 2006.

[2] C. Kulkarni, K. Pang-Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer, "Peer and self assessment in massive online classes," Stanford University, Tech. Rep., 2013.

[3] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and 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, and K. Ramchandran, "A case for ordinal peer-evaluation in MOOCs," in *NIPS Workshop on Data Driven Education*, 2013.

[5] K. Raman and 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, and 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, and A. Bahamonde, "A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments," *Knowledge-Based Systems*, 2015.

[8] K. Raman and T. Joachims, "Bayesian ordinal peer grading," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, ser. L@S '15. New York, NY, USA: ACM, 2015, pp. 149–156.

[9] A. Bahamonde, G. F. Bayón, J. Díez, J. R. Quevedo, O. Luaces, J. J. del Coz, J. Alonso, and 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.

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

[11] W. Barnett, "The modern theory of consumer behavior: Ordinal or cardinal?" *The Quarterly Journal of Austrian Economics*, vol. 6, no. 1, pp. 41–65, 2003.

[12] J. J. del Coz, G. F. Bayón, J. Díez, O. Luaces, A. Bahamonde, and 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] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais, "Here or There. Preference Judgments for Relevance," in *Advances in Information Retrieval*. Springer, 2008, pp. 16–27.

[14] J. A. Krosnick, "Survey research," *Annual review of psychology*, vol. 50, no. 1, pp. 537–567, 1999.

[15] R. Herbrich, T. Graepel, and 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] V. Aggarwal, S. Srikant, and 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.

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

[18] D. Pérez-Marín, I. Pascual-Nieto, and P. Rodríguez, "Computer-assisted assessment of free-text answers," *The Knowledge Engineering Review*, vol. 24, pp. 353 – 374, 2009.

[19] F. Rodrigues and P. Oliveira, "A system for formative assessment and monitoring of students' progress," *Computers & Education*, vol. 76, no. 0, pp. 30 – 41, 2014.

[20] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[21] K. Papineni, S. Roukos, T. Ward, and 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.

[22] F. Noorbehbahani and A. Kardan, "The automatic assessment of free text answers using a modified BLEU algorithm," *Computers & Education*, vol. 56, no. 2, pp. 337 – 345, 2011.

[23] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[24] P. Thomas, D. Haley, A. deRoeck, and 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.

[25] R. Blake and O. Gutierrez, "A semantic analysis approach for assessing professionalism using free-form text entered online," *Computers in Human Behavior*, vol. 27, no. 6, pp. 2249 – 2262, 2011.

[26] D. Pérez, A. M. Gliozzo, C. Strapparava, E. Alfonseca, P. Rodríguez, and 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.

[27] C. Gütl, "Moving towards a fully automatic knowledge assessment tool," *International Journal of Emerging Technologies in Learning*, vol. 3, no. 1, 2008.

[28] H.-C. Wang, C.-Y. Chang, and T.-Y. Li, "Assessing creative problem-solving with automated text grading," *Computers & Education*, vol. 51, no. 4, pp. 1450 – 1466, 2008.

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