A peer assessment method to provide feedback, consistent grading and reduce students' burden in massive teaching settings

Oscar Luaces[a,*], Jorge Díez[a], Antonio Bahamonde[a]

[a] *Artificial Intelligence Center*
University of Oviedo
33204 Gijón, Spain

Abstract

To grade open-response answers in a massive course is an important task that cannot be handled without the assistance of an intelligent system able to extend the abilities of experts. A peer assessment method may be used for this. The students who wrote the answers also play the role of graders for a reduced set of answers provided by other students. The grades thus obtained should be aggregated to provide a reasonable overall grade for each answer. However, these systems present two clear disadvantages for students: they increase their already heavy workload, and the grades that students finally receive lack feedback explaining the reasons for their scores. The contribution of this paper comprises a proposal to overcome these shortcomings. The students acting as graders are asked to evaluate a number of different aspects. One of them is the overall grade, but there are other annotations that can be included to explain the overall grade. Moreover, we represent the responses given by the students (text documents) as the inputs in a learning task, in which the outputs are the aspects to be assessed (labels with an ordinal level). Our proposal is to learn all these labels at once employing a multitask approach that uses matrix factorization. The method presented in this paper shows that peer assessment can provide feedback and can additionally be extended to grade the responses of students not involved in the peer assessment loop, thus significantly reducing the burden on students. We present the details of the method, as well as a number of experiments carried out using three data sets obtained from courses belonging to different fields at our university.

Keywords: Peer Assessment, Students' Burden Reduction, Feedback, Factorization, Preference Learning

*Corresponding author: Tel: +34 985 182 028
Email addresses: oluaces@uniovi.es (Oscar Luaces), jdiez@uniovi.es (Jorge Díez), abahamonde@uniovi.es (Antonio Bahamonde)

1. Introduction

The assessment of open-response assignments is frequently a problem. This is the case in massive courses like MOOCs or even when there are a lot of assignments during a course. One of the options to overcome this problem is to avoid open-response in favor of multiple-choice questions. However, this significantly reduces the communication between students and instructors that may involve handling different forms of data, including computer programs, video, audio, and written texts. The alternative is for the students that wrote the answers to also play a role in the assessment. Peer assessment has been explored as an efficient procedure to deal with this problem; see for instance (Kulkarni et al., 2015; Piech et al., 2013; Raman and Joachims, 2014, 2015; Sadler and Good, 2006; Shah et al., 2013; Labutov and Studer, 2016; Díez et al., 2013; Luaces et al., 2015a,b, 2017; Formanek et al., 2017). It has been acknowledged as an activity that enhances student learning in Sun et al. (2015).

However, peer assessment has a number of flaws that should be addressed in order to be deployed more extensively. Firstly, peer assessment may considerably increase the burden on students. Second, the quality of the feedback received by students should be improved (Gielen et al., 2010; Liu and Carless, 2006; Tseng and Tsai, 2007; Hovardas et al., 2014). In addition to a grade, students should obtain some annotations pointing to the weak and strong aspects of their answers. Lu and Law (2012) present an interesting analysis on the effects of providing feedback both for the students being assessed, and for the students acting as graders. The authors conclude that feedback is clearly beneficial and they also point out some aspects that must be considered in order to design good peer assessment processes.

Finally, the fact that each assignment is evaluated by different peer assessors gives rise to the problem of reaching a consensus to summarize those different opinions into a single grade and a feedback. This is not trivial, as we will explain in Section 2.

Table 1 shows some of the approaches proposed to deal with the evaluation of assignments in a peer assessment context. Most of them are content-based, i.e. the content of the answers (words) is represented and used to build the assessment model. Note also that none of the approaches shown in the table provides any kind of feedback other than the final grade).

In this paper we explore a method to address the mentioned issues in peer assessment when open-responses are written documents. To improve the feedback from an automated perspective, we propose to use a set of labels or annotations that may be attached to answers with a level. These labels should cover the explanations that a student could obtain from a personalized assessment given by a professional instructor. We tested this proposal in three courses at our university belonging to different fields: Law and Economy. Instructors could easily express the possibilities of annotations in terms of labels with levels. On the other hand, the students were able to understand the assessment task with annotations effortlessly.

The output of peer assessments is a data set that must somehow be filtered

Table 1: Several approaches to deal with automatic grading in peer assessment. Column C-b indicates whether the method is content-based. None of the approaches produce feedback to the students other than the final grade

Reference	C-b	Summary
Shah et al. (2014)	☐	Very abstract proposals which include dimensionality reduction and using features involved in the assessment.
Rodrigues and Oliveira (2014)	☒	Answers, represented using the vector space model (VSM) (Salton et al., 1975), are matched with a reference text (correct answer) using the cosine as a similarity metric. This method includes semantic analysis, whereby two words are considered to be similar if they are related in the WordNet semantic network.
Noorbehbahani and Kardan (2011)	☒	Answers are matched with a reference text using a modified version of the BLEU metric (Papineni et al., 2002).
Thomas et al. (2004)	☒	This approach employs *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990), which uses the singular value decomposition (SVD) of the term-document matrix (matrix of answers codified using the vector space model (VSM) representation) to project it into a smaller dimensional space.
Guetl (2007)	☒	The *CarmelTC* algorithm implements a naïve Bayes classifier which requires a reduced set of answers to be graded by an instructor and then processed by an ordinal classifier that learns from the supervised data set.
Wang et al. (2008)	☒	This approach uses a support vector machine (SVM) to evaluate creative problem-solving from open-ended responses.

to aggregate or reconcile the grades received by one answer from several students acting as graders without experience in this task. This is usually tackled using Machine Learning methods. In the experiments reported at the end of the paper, we prove that models learned to aggregate grades can be used to ease the academic workload of students.

The idea is to extend the assessment model to answers not involved in peer assessment in any way. For this purpose, we use a content-based approach similar to those employed in Recommender Systems. In this context, contents are documents (the students' answers) that can be represented using a *bag of words* representation. Several approaches have been proposed to overcome the limitations of this representation paradigm; see for instance (Deerwester et al., 1990). In this paper, we propose a matrix factorization method to learn how to grade that includes a method to arrange the answers of students in a metric space according to their grades.

Notice that we have to learn to grade each of the aspects of the answer that need to be considered: the overall grade, and the level of each of the labels or annotations for feedback. We present a multitask (Caruana, 1997) method to simultaneously learn all the aspects to be assessed, and we show in the experiments that, in fact, there is an inductive transfer that improves the whole Machine Learning process.

The contributions of this article can be summarized as follows. The eval-

Criteria	Levels										
The answer contains misspellings	○ many		○ some		○ few		● none				
Quality of the composition	○ bad		○ improvable		○ acceptable		● good				
Short-term financial analysis	○ deficient		○ insufficient		○ sufficient		○ good		● excellent		
Long-term financial analysis	○ deficient		○ insufficient		○ sufficient		● good		○ excellent		
Economic analysis	○ deficient		● insufficient		○ sufficient		○ good		○ excellent		
Overall grade	○ 0	○ 1	○ 2	○ 3	○ 4	● 5	○ 6	○ 7	○ 8	○ 9	○ 10

Figure 1: Template used to annotate the assessment of the answers in the assignment of *Accounting Information*, see Section 4.1

uation of open response assignments is a task that must be carried out by an expert. However, when there exists a large number of assignments, which will take the instructor(s) a prohibitive time to assess, peer assessment is commonly used. The inherent subjectivity in the assessment, as well as the fact that each assignment will receive different marks from different peer assessors, poses the need to aggregate those scores. Computing the average is risky, since we have only a few assessments per assignment, so we need to use an smarter approach. Thus, the help of an intelligent system capable of performing this task is needed.

Generally speaking, an intelligent system is a piece of software able to perform a task which requires some kind of intelligent behaviour. In this context, our proposed method is able to generalize the criteria of the peer assessors (graders), going beyond averaging their scores, and getting rid of their subjectivity in the assessment.

However, peer assessment entails an added burden to the already large volume of work required to the students. The first contribution of this paper is a method that does not require all students to participate in the peer assessment. The second is that our method provides feedback to students in the form of a summarized explanation of the final grade, something which is not done automatically by other peer assessment methods.

The paper is organized as follows. First we explain the whole process as it is seen by students and instructors and then introduce the insight behind the approach presented here. The following section is devoted to presenting the formal setting. We then report the experiments conducted to evaluate the approach presented in the paper. We end with the conclusions of this research.

2. Overall Description of the Method

After submitting their answers to an open-response assignment, students are required to grade a group of anonymized answers by other students. The

assessments must be carried out using a template like the one shown in Figure 1, and following the rules detailed in a *rubric*. Notice that the template presents a set of graded annotations or labels that, together with the overall grade, will form the feedback for the students who wrote the answers. Thus, the feedback provided to the student with the scores shown in Figure 1 can be read as:

> *You have got 5 points out of 10 (overall grade) because you made an economic analysis which was insufficient, although the long-term financial analysis was good and the short-term analysis was excellent. The quality of your composition was good, with no misspellings.*

The rubric should include the *correct* answer when this is not clear for all students; this was the case of the assignment of *Constitutional Law* used in the experiments reported in Section 4. However, at other times, for instance in the course entitled *Spanish Economy*, the rubric indicates what elements would contain a good answer, and the grade is to a certain extent a subjective opinion of the grader.

The labels and the rubric must be provided by the instructor considering that they serve to organize the process of assessment. They should aim to achieve uniform assessment criteria.

The output of peer assessment is a 3-dimensional matrix like that in Figure 2. In the figure, the answers are represented in columns, the graders in rows, and the labels to be graded in pages. Most of the components of this matrix are usually empty.

We assume that there is an unknown relationship between the grades given to the labels of the answers. So if we find a pattern in the grades of some of these labels, we hope to use them explicitly as an inductive transfer to learn how to make an assessment of all the labels of all the answers by all the graders. In other words, we try to complete the assessment matrix with grades *consistent* with those we have available, as shown in Figure 2.

The *consistency* of grades with the original assessment matrix is established in term of orderings. The aim is to have a ranking of answers as similar as possible to the partial rankings provided by graders. Thus, we will not to use regression. The main reason is that graders are not professionals. Therefore, if a grader assigns 9 points to an answer $\boldsymbol{x}$ and 4 points to $\boldsymbol{y}$, we are only using the fact that $\boldsymbol{x}$ is *preferable to* $\boldsymbol{y}$. This is the *ordinal* point of view. If we were trying to learn how to predict exactly 9 points for $\boldsymbol{x}$ and 4 for $\boldsymbol{y}$ (the regression approach), then we would had adopt the *cardinal* point of view.

There are many reasons in favor of the ordinal approach, not only in assessment, but in general when we are interested in learning preferences in contexts like information retrieval or marketing studies (Bahamonde et al., 2004; Joachims, 2002; Luaces et al., 2015c).

Once we have a complete estimation of the assessment matrix, we compute the average of all the grades assigned to each answer in each label, including the overall grade. These values will be the *grades* given by the model just learned from peer assessment data. However, these grades are just a tool to order the answers.

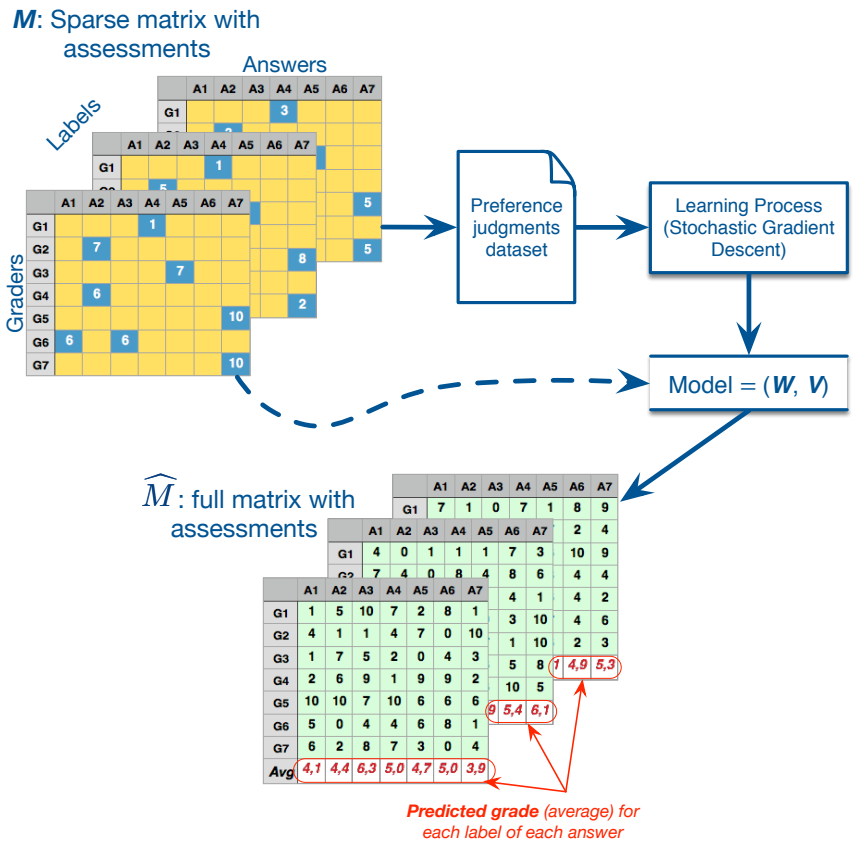

Figure 2: The process starts out from a sparse assessment matrix and provides a full matrix after learning from the available data

Sometimes, these rankings (one for each label) are enough to complete the assessment process. If this is not the case (as happens in Spanish universities), we need to compute a grade. We then transform ranking positions into grades following the same distribution as those given by the students acting as graders. In this respect, we take into account the grades given by the graders. But let us emphasize here that this final step is just a translation from the language of percentiles to that of grades.

3. Formal Settings

In this section we present the formal details of our proposal. Firstly, we explain a multitask approach, which learns a predictive model taking into account all the available data, i.e. all the labels simultaneously, including the overall mark. Then, we present a simplification in which a different model is learned for each label. In the end of the section we suggest how the rankings obtained by any of the previous methods can be transformed into an absolute grade, just in case it was necessary.

3.1. Multitask approach

Let $\mathcal{G}$ be a set of *graders*, and $\mathcal{A}$ a set of answers for an assignment. Graders are asked to assign a grade for a set of aspects of the assignment represented by labels in $\mathcal{L}$. After the assessment, we have an *assessment matrix*,

$$\boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}). \tag{1}$$

The values of this matrix are grades given by a grader $\boldsymbol{g} \in \mathcal{G}$ for answer $\boldsymbol{a} \in \mathcal{A}$ with respect to a label $\boldsymbol{l} \in \mathcal{L}$. Typically, one of the labels stands for the *overall grade*, but formally this is only another label to be assessed. The remaining labels will be understood as feedback given to the students who wrote the answers to explain the final grade.

Not all components of $\boldsymbol{M}$ will have values. In fact, these type of matrices are often quite sparse. The reason for this is that each grader $\boldsymbol{g}$ is asked to evaluate only a few answers $\mathcal{A}_{\boldsymbol{g}} \in \mathcal{A}$. As mentioned above, the first step toward an assessment is to fill the matrix according to the available values. For this purpose, we start out from a set of *preference judgments* (Bahamonde et al., 2004; Joachims, 2002; Luaces et al., 2015c),

$$\mathcal{D} = \{(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^b, \boldsymbol{a}^w) : \ \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^b) > \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^w)\}, \tag{2}$$

where $\boldsymbol{g} \in \mathcal{G}$, $\boldsymbol{l} \in \mathcal{L}$, and $\boldsymbol{a}^b, \boldsymbol{a}^w \in \mathcal{A}_{\boldsymbol{g}}$. The intended meaning is to record that for $\boldsymbol{g}$, for label $\boldsymbol{l}$, the answer $\boldsymbol{a}^b$ deserves a higher grade than the answer $\boldsymbol{a}^w$. In this way, we overcome the actual grades, but retain the ordinal preferences of the graders. Nevertheless, at the end of the process we will take into account the distribution of grades given for each label, just in case we need to transform the final ranking into absolute scores.

We use a vectorial representation to handle answers, labels, and graders. We thus use a *bag of words* approach for the answers to explicitly consider the contents of the answers in the assessment method. This requires first computing the *corpus* of all words used in all answers in $\mathcal{A}$. Then, each answer can be codified by a binary vector indexed by the corpus: the components corresponding to a word that appears in the answer will have a value of 1, while the rest will have a 0. This is a straightforward approach to implement a content-based system that can be eventually replaced by other, most advanced, representation methods, like those proposed by Le and Mikolov (2014) or by Cho et al. (2014).

On the other hand, graders and labels will be represented using *one-hot* codification. The i-th element will be codified by a vector whose only nonzero value will be the i-th component. Then, to consider grader $\boldsymbol{g}$ and label $\boldsymbol{l}$ at the same time, we use the *direct sum* (concatenation) of their vectorial representations, $(\boldsymbol{g} \oplus \boldsymbol{l})$.

All vectors involved in the assessment process will be projected (*embedded*) in a common Euclidean space, $\mathbb{R}^k$,

$$\mathbb{R}^{|\mathcal{G}|+|\mathcal{L}|} \mapsto \mathbb{R}^k, \qquad (\boldsymbol{g} \oplus \boldsymbol{l}) \mapsto \boldsymbol{W}(\boldsymbol{g} \oplus \boldsymbol{l}), \tag{3}$$

$$\mathbb{R}^{|\operatorname{corpus}(\mathcal{A})|} \mapsto \mathbb{R}^k, \qquad \boldsymbol{a} \mapsto \boldsymbol{V}\boldsymbol{a}. \tag{4}$$

Notice that the input of projections depends on the size of the corpus and on the number of labels and graders, while $\mathbb{R}^k$ has an arbitrary dimension, k. We normally use a lower dimension than that of input spaces. The idea is to reduce the noise of the data.

In this context, we define a *full* assessment matrix, $\widehat{\boldsymbol{M}}$, to estimate the grade for a label $\boldsymbol{l}$ given to an answer $\boldsymbol{a}$ according to grader $\boldsymbol{g}$, using the inner product of the projections in $\mathbb{R}^k$ as follows:

$$\widehat{\boldsymbol{M}}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}) = \langle \boldsymbol{W}(\boldsymbol{g} \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a} \rangle = (\boldsymbol{g} \oplus \boldsymbol{l})^T \boldsymbol{W}^T \boldsymbol{V}\boldsymbol{a}. \tag{5}$$

In this equation, the matrices $\boldsymbol{W}^T$ and $\boldsymbol{V}$ are *factors* of a matrix of weights for the products of the components of $(\boldsymbol{g} \oplus \boldsymbol{l})$ and $\boldsymbol{a}$. For this reason, this approach is called *matrix factorization*.

Finally, the grade for aspect $\boldsymbol{l}$ of answer $\boldsymbol{a}$ is defined by the average of grades given by all graders using the estimations of $\widehat{\boldsymbol{M}}$; see Figure 2. In symbols,

$$f(\boldsymbol{l}, \boldsymbol{a}) = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{g} \in \mathcal{G}} \widehat{\boldsymbol{M}}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}) = \langle \boldsymbol{W}(\bar{\boldsymbol{g}} \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a} \rangle, \tag{6}$$

where $\bar{\boldsymbol{g}}$ stands for the *average grader*,

$$\bar{\boldsymbol{g}} = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{g} \in \mathcal{G}} \boldsymbol{g}. \tag{7}$$

The coherence of the assessment matrix, $\boldsymbol{M}$, and its estimation, $\widehat{\boldsymbol{M}}$, is measured in terms of differences in the orderings of the answers. The aim is for the

orderings induced by the estimated grades (of the average grader and each of the graders) to be as similar as possible to the ordering given by each grader. Then we search for the best matrices, $\boldsymbol{W}$ and $\boldsymbol{V}$. The formalization of our multitask approach is that both parameters, $\boldsymbol{W}$ and $\boldsymbol{V}$, are the same for all labels.

To measure the similarity of the orderings, we use a maximum margin approach. We seek to reduce the number of swapped pairs in the orderings. The optimization problem considering all labels at the same time can be set to minimize the following loss function:

$$\begin{aligned}\mathrm{err}(\boldsymbol{W},\boldsymbol{V}) = \sum_{(\boldsymbol{g},\boldsymbol{l},\boldsymbol{a}^b,\boldsymbol{a}^w)\in\mathcal{D}} \max\Big\{0, 1-\langle \boldsymbol{W}((\bar{\boldsymbol{g}}+\boldsymbol{g})\oplus\boldsymbol{l}), \boldsymbol{V}\boldsymbol{a}^b\rangle + \\ \langle \boldsymbol{W}((\bar{\boldsymbol{g}}+\boldsymbol{g})\oplus\boldsymbol{l}), \boldsymbol{V}\boldsymbol{a}^w\rangle\Big\}.\end{aligned} \tag{8}$$

To solve this optimization problem, we use a Stochastic Gradient Descent (*SGD*) that finds the optimum parameters of the model, i.e. the matrices $\boldsymbol{W}$ and $\boldsymbol{V}$. Appendix A gives more mathematical details about the gradient descent procedure.

3.2. Binary Relevance (BR) approach

The straightforward baseline for the multitask approach is to learn one model for each label. In our case, to learn matrices $\boldsymbol{W_l}$ and $\boldsymbol{V_l}$ for each label $\boldsymbol{l}$. For this purpose, we need to focus only on those grades involving one label $\boldsymbol{l}$,

$$\mathcal{D}_{\boldsymbol{l}} = \{(\boldsymbol{g},\boldsymbol{a}^b,\boldsymbol{a}^w) : (\boldsymbol{g},\boldsymbol{l},\boldsymbol{a}^b,\boldsymbol{a}^w)\in\mathcal{D}\}. \tag{9}$$

Using the same approach as that presented above for the multitask approach, we estimate the grades for an answer $\boldsymbol{a}$ and a label $\boldsymbol{l}$ using a particular function

$$f_l(\boldsymbol{a}) = \langle \boldsymbol{W}_l\,\bar{\boldsymbol{g}}, \boldsymbol{V}_l\,\boldsymbol{a}\rangle. \tag{10}$$

In the following, we will refer to this simple approach as *Binary Relevance* (BR) using a terminology borrowed from *multilabel* classification.

Figure 3 depicts an artificial example to highlight the differences between the multitask and BR approaches from a geometrical point of view. Let us recall that, in the multitask approach, whose scoring function takes the form expressed in (6), vector $\boldsymbol{W}(\bar{\boldsymbol{g}}\oplus\boldsymbol{l})$ defines an assessment hyperplane in $\mathbb{R}^k$ for label $\boldsymbol{l}$, such that the model predicts that the score (grade) given by the average grader to answer $\boldsymbol{a}_i$ regarding aspect (label) $\boldsymbol{l}$ is proportional to the distance from its representation in $\mathbb{R}^k$, $\boldsymbol{V}\boldsymbol{a}_i$, to the assessment hyperplane. The parameters of the multitask model, $\boldsymbol{W}$ and $\boldsymbol{V}$, are learned simultaneously for all labels.

In contrast, for the BR approach, the problem is split into as many problems as the number of labels to be assessed, so there will be a different model, i.e. a different pair of matrices $\boldsymbol{W}_l$ and $\boldsymbol{V}_l$ for each label. In other words, the BR approach consists in decomposing the original learning problem into as many independent subproblems as the number of labels to be assessed.

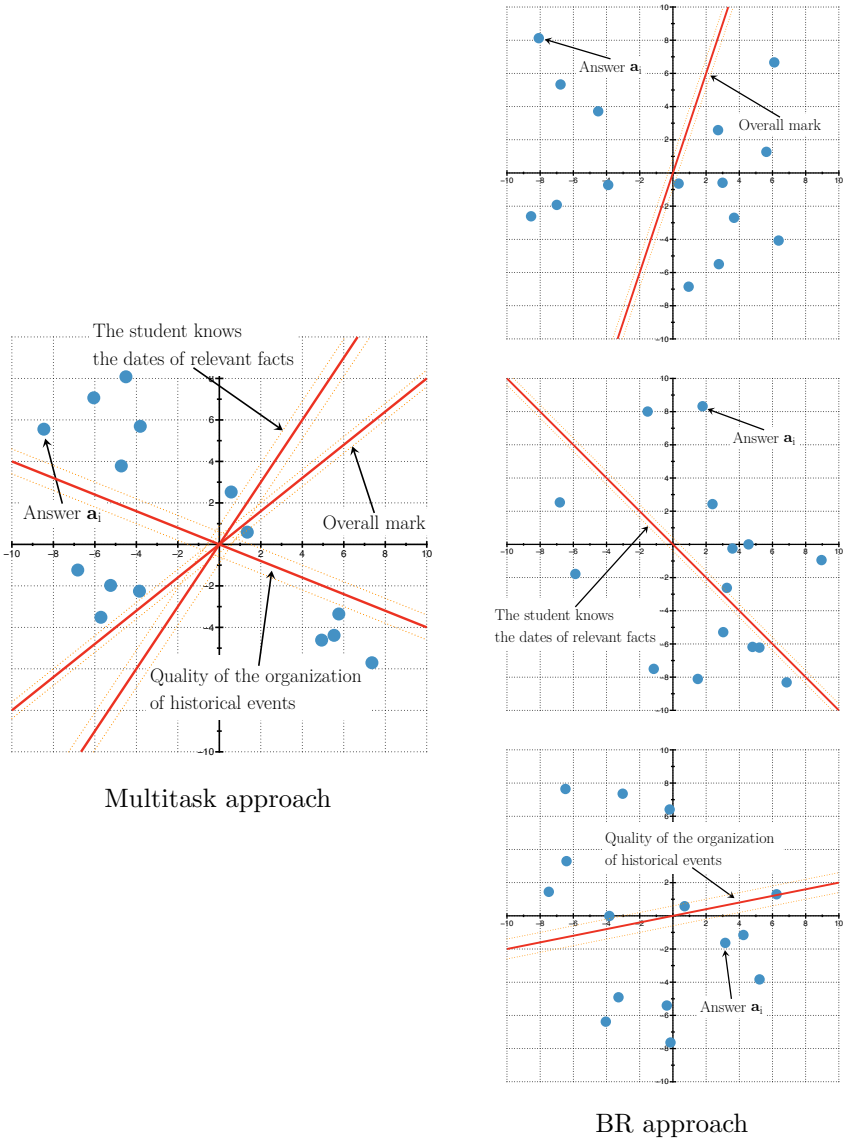

Figure 3: Geometrical interpretation of the multitask approach (left) vs. the binary relevance approach (right) for a hypothetical History assignment. The blue points represent answers and the red lines are the hyperplanes defined by the graders and the assessment labels. In this example, the labels to be assessed are *the quality of the organization of historical events* in the answer, *the student's knowledge regarding relevant facts* in a given historical period, and finally, the *overall mark*. In the multitask approach, the location of answers ($\boldsymbol{V}\boldsymbol{a}$) and hyperplanes ($\boldsymbol{W}$ $(\bar{\boldsymbol{g}} \oplus \boldsymbol{l})$) is learned to cope with all labels at the same time, aiming to take advantage of an inductive transfer between them, while in the BR approach, there is a model for each label, i.e. a different pair of matrices $\boldsymbol{W}_l$ and $\boldsymbol{V}_l$, yielding in this example 3 different graphs

Table 2: Characteristics of the datasets. The total number of students participating in each peer assessment event is indicated in *Number of answers*. Note that only in Constitutional Law did all the students also act as graders

	Account Information	Constitutional Law	Spanish Economy
Number of answers	119	66	111
Number of graders	112	66	108
Number of assessments	1120	660	1065
Sparsity of $\boldsymbol{M}$ (% empty)	92.09	84.85	91.36
Average number of grades per answer	9.41 ± 0.71	10 ± 0	9.59 ± 0.67
Average number of grades per grader	10 ± 0	10 ± 0	9.86 ± 0.99

Thus, the work entrusted to the parameters of the model ($\{\boldsymbol{W}, \boldsymbol{V}\}$ in the multitask approach, and several pairs of $\{\boldsymbol{W}_l, \boldsymbol{V}_l\}$ in the BR approach) is to place hyperplanes and answers in $\mathbb{R}^k$ in such a way that the distances to the corresponding hyperplanes are coherent with the ordering of graders. The multitask approach places the representations of answers (blue points in the figure) in the same position for all labels, while Binary Relevance starts from scratch for each label and hence the clues given by one label cannot be used in any way to place the projections of answers for another label.

3.3. Transforming the ranking into a grade

After learning the matrices $\boldsymbol{W}$ and $\boldsymbol{V}$, using the function f (6), for each label, we have a ranking of answers from best to worse. If we need to transform this ranking into grades, our proposal is to do so by trying to reproduce the same distribution of grades that we collected from the students. Notice that this is only a translation that has no effect on the ranking of answers learned in the multitask described above.

Of course, we may follow an analogous process for BR to obtain grades for each answer in each label.

4. Experimental Results

In this section we report a number of experiments performed to test the goodness of the method presented in this paper. First, we introduce the datasets used, then the evaluation method, and finally the scores obtained.

4.1. Datasets

The datasets used in the experiments were gathered from three courses belonging to different fields at the *University of Oviedo*: Accounting Information, Constitutional Law, and Spanish Economy.

To collect the data, we used a Moodle (moodle.org) installation on one of our servers. This platform has a tool called *workshop* that provides the infrastructure required for peer assessment. The final grade is computed in this tool

by averaging the grade received by each answer, so we replaced this step by our method.

The double-blind assessment also guaranteed that no student graded her or his own answer. Each student received 10 answers to grade.

Table 2 shows the basic characteristics of the datasets. Note that around 90% of the components of the assessment matrices $\boldsymbol{M}$ (1) are empty.

On the other hand, Table 3 shows additional characteristics of the datasets. The first column (*#PJ*) reports the number of *Preference Judgments*, i.e. the size of the corresponding dataset $\mathcal{D}_{\boldsymbol{l}}$ (9). Notice that for each label and each grader we record only those pairs of answers with different grades; the pairs with the same grade do not give rise to any element in $\mathcal{D}_{\boldsymbol{l}}$. This is the reason why the number of preference judgments may be different for different labels. Recall that the multitask approach deals with the join of $\mathcal{D}_{\boldsymbol{l}}$ for all $\boldsymbol{l}$ in a single $\mathcal{D}$ (2).

The second column (*Discrepancies*) gives the percentage of contradictory preferences with the majority of opinions. For instance, if, for a label $\boldsymbol{l}$, 3 graders think that answer $\boldsymbol{x}$ is better than $\boldsymbol{y}$ and another 2 graders think the opposite, we count 2 discrepancies. Thus, the percentage of discrepancies is a lower error bound for any classification function.

Finally, the last column in Table 3 details the set of labels used in the assignments whose data were used in the experiments. The original labels were written in Spanish, so here we give a translation. The number in parentheses is the same as that used in Figure 4. This picture represents the distribution of grades given by graders for all labels including the overall grade in the rightmost graph in each row.

4.2. Evaluation Method

To evaluate the performance of the multitask approach presented in this paper, we conducted some train/test experiments with the datasets described above. We compared the performance of multitask versus BR (Section 3.2). To split the datasets, we first separated a set of students and made a training set with only the preference judgments involving this subset of students, either as graders or as those who gave the answers. The remaining available preference judgments were then considered as the test set. In other words, starting from the sparse matrix, $\boldsymbol{M}$, we extract several matrices, $\boldsymbol{M}_s$, built by randomly selecting s rows and the corresponding columns from $\boldsymbol{M}$, leaving the rest for testing. The size of the set of students selected was $s = \{25, 50, 75, 100\}$, except in the case of the dataset from Constitutional Law as we only had 66 students and therefore we only considered training sets of students of size 25 and 50.

The performance measure was a simple 0/1 classification error in the test. The errors are those ordered pairs of answers in the test sets that were not ordered in the same way by the function shown in (6) learned by the multitask approach. Table 4 shows the percentage of errors computed averaging 10 repetitions.

During training, the SGD algorithm uses some parameters that must be set in order to ensure the best performance. For this purpose, we made a grid search

Table 3: Detailed description of grades and labels for each dataset. The #PJ column indicates the number of pairs of comparisons, i.e. the pairs of assignments with a different grade in the corresponding label, given by the same grader. The Discrep. (%) column indicates the percentage of pairs ordered differently depending on the grader (discrepancies between graders for a given pair of assignments). The Labels column contains the aspects that had to be assessed in each assignment and that can be used as feedback to students when they receive their marks (see Section 2)

Accounting Information

#PJ	Discrep. (%)	Labels
1603	3.74	(1) The answer contains misspellings
3068	5.05	(2) Quality of the composition
3043	4.24	(3) Short-term financial analysis
3187	4.61	(4) Long-term financial analysis
3455	4.08	(5) Economic analysis
4233	5.20	Overall grade

Constitutional Law

#PJ	Discrep. (%)	Labels
570	2.98	(1) The answer contains misspellings
1273	6.44	(2) Quality of the composition
1172	4.95	(3) Line of arguments
378	2.91	(4) Quotes the relevant papers
171	0.00	(5) Does not know what a motion of censure is
218	2.29	(6) Does not know what a motion of non-confidence is
369	1.90	(7) Does not know the duties of the King
184	0.00	(8) Does not know how the President is appointed
112	2.68	(9) Does not know the duties of the President
2158	9.73	Overall grade

Spanish Economy

#PJ	Discrep. (%)	Labels
2318	5.95	(1) Ability to understand and describe the core economic processes of each of the stages of evolution of the Spanish Economy
2331	5.19	(2) Ability to distinguish the phases of convergence and divergence of the Spanish Economy with respect to the European economy
2329	5.84	(3) Ability to show the overall balance of the evolution of the Spanish Economy with its main achievements and limitations
2544	6.29	(4) Ability to reasonably explain the salient features, events and consequences of the recent economic crisis and the dilemmas posed to economic policies
2735	5.45	(5) Quotes the relevant references and uses well-reasoned personal judgments
2648	6.50	(6) The arguments are well organized and clear. The answer shows the capacity to summarize and uses the right economic terms
3736	8.00	Overall grade

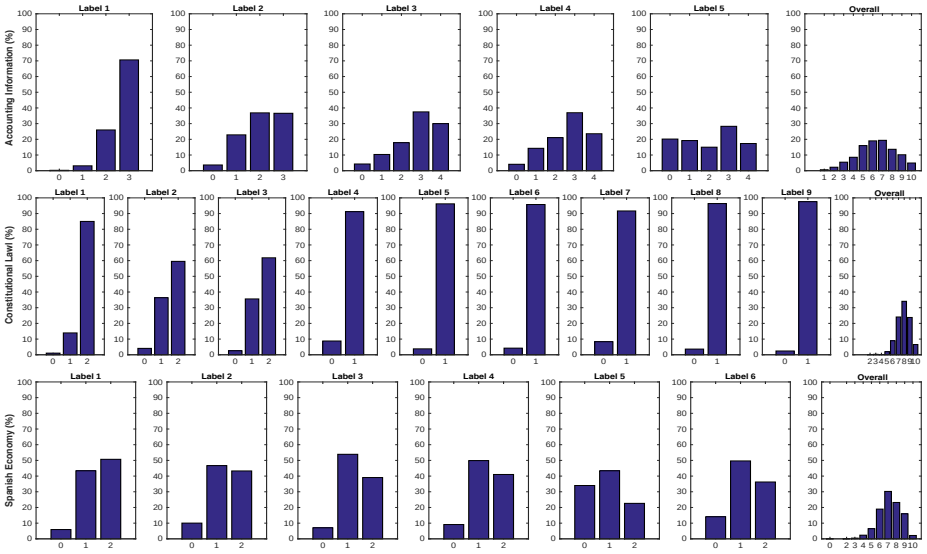

Figure 4: Distributions of the scores given by graders for all labels in the three peer assessment experiments. The first row contains the distribution for the labels assessed in Account Information, the second row is for Constitutional Law, and the third is for Spanish Economy. These labels can be found in Table 3. The X axis of each graph shows the number of different scores that the corresponding label can receive (for instance, the overall mark in any of the three assignments is a value in the range $[0, 10]$)

using only training sets to find the most promising combination. We conducted a cross-validation experiment with 2 folds and 5 repetitions using all possible combinations of values of k, ν and γ_s (A.2):

$$k \in \{2, 10, 20, 50, 100\},$$
$$\nu \in \{10^e : e = -4, \ldots + 2\},$$
$$\gamma_s \in \{10^e : e = -4, \ldots, -1\}.$$

We then selected the best combination to perform the corresponding train/test experiment.

4.3. Results

Table 4 reports the scores obtained by the multitask and BR approaches. The last row of each table shows the weighted average of the scores of all labels for each train/test; the weights are the number of test elements.

The best (weighted average) scores are highlighted in bold. It can be seen that the multitask approach outperforms the BR approach in most cases: multitask is better 8 out of 10 times.

As expected, the error decreases as the size of the training set increases. The scores obtained in the biggest sets are slightly higher than those reported in the

Table 4: Percentage of errors. There is a table for each independent dataset (assignment). Errors are the percentage of incorrectly ordered pairs of assignments when using only 25, 50, 75 or 100 graders/assignments for training, and the rest for testing

Accounting Information

	Multitask				BR			
	25	50	75	100	25	50	75	100
L1	43.29	45.79	48.10	44.26	49.47	47.74	46.08	44.00
L2	36.86	31.24	36.14	28.75	45.22	32.41	32.78	27.74
L3	48.24	30.96	30.44	23.72	48.26	33.01	27.88	23.89
L4	41.91	35.86	32.99	29.42	47.24	36.27	32.08	30.60
L5	40.50	30.27	31.57	25.53	44.57	33.57	29.90	26.34
Overall	40.22	29.60	28.21	25.30	45.85	30.09	26.67	25.30
weighted	**41.58**	**32.68**	32.99	**27.96**	46.45	34.17	**31.03**	28.15

Constitutional Law

	Multitask		BR	
	25	50	25	50
L1	43.24	36.82	39.71	38.32
L2	34.07	33.90	35.73	37.45
L3	28.97	30.41	30.76	34.33
L4	35.80	23.63	46.56	27.81
L5	43.01	36.81	51.70	36.39
L6	31.39	33.08	43.56	27.05
L7	42.28	18.42	40.73	23.51
L8	36.47	26.43	42.88	30.86
L9	26.87	31.43	44.34	30.86
Overall	34.95	36.40	36.92	35.62
weighted	**34.88**	**33.04**	37.46	34.68

Spanish Economy

	Multitask				BR			
	25	50	75	100	25	50	75	100
L1	54.94	40.65	39.71	32.13	49.57	47.98	43.49	38.03
L2	52.46	40.72	36.75	37.07	43.24	45.41	40.85	39.66
L3	51.54	40.72	36.69	30.78	47.08	44.65	41.86	37.51
L4	50.44	42.61	40.60	31.51	48.42	46.17	43.98	36.31
L5	42.62	36.23	35.92	31.25	45.09	41.04	37.20	35.81
L6	49.14	43.04	40.53	40.38	49.26	48.33	43.01	39.58
Overall	49.06	40.63	38.40	33.61	48.23	46.41	38.92	37.96
weighted	49.78	**40.62**	**38.38**	**33.79**	**47.34**	45.72	41.11	37.82

papers by Raman and Joachims (2014); Luaces et al. (2015a,b, 2017). The main reason for this is that the aforementioned papers report the resubstitution error; the comparison was established comparing the discrepancies between professional instructors and the model learned.

In this paper, we present a collection of train/test experiments that, on the one hand, provide a support to launch assessment tools where only a part of the students will be required to grade their peers. On the other hand, the experiments reported here back the hypothesis that assessment can be smoothly learned like other learning tasks.

5. Conclusions

In this paper, we address two important issues in order to increase the quality of the peer assessment of written open-responses: the need to provide useful feedback to students, and to relieve their workload. The proposal requires graders to assess a number of annotations or labels about the answer that they are assessing. The overall grade is another label in this context. We have presented a method that uses a multitask approach to search for grading patterns in all labels at the same time.

Multitask leverages the accuracy of a baseline that successively focuses on each label separately. Thus, the assessments provided by students can be aggregated in a list of graded labels that informs their peers of their overall grade, as well as providing a number of reasons explaining weak and strong points in their answers.

On the other hand, models learned using the multitask approach can be extended to answers not at all involved in peer assessment. The consequence is that a part of the students can be relieved of the assessment task, thereby reducing the burden on students in processes of this kind.

The paper thus presents an intelligent system that allows implementing the assessment of open-response assignments in massive courses. Our proposal presents two fundamental contributions with respect to the existing literature. On the one hand, it considerably reduces the student workload. In fact, when students are required to assess the work of others, they must spend a lot of time on this task. In our proposal, by explicitly using the texts of the answers, the intelligent system is able to do the work of an expert evaluator: it generalizes what has been learned from a part of the students to grade the rest. On the other hand, the system presented here is capable of offering an expert explanation to students, in addition to an overall grade to their answers. This explanation constitutes a feedback that undoubtedly contributes to improving teaching.

The goodness of the approach presented in this paper was checked on three datasets collected from courses at our university (Accounting Information, Constitutional Law, and Spanish Economy), yielding quite successful accuracy scores. Therefore, we would like to stress that this research proved that it is feasible to deploy sophisticated assessment methods in fields far-removed from Computer Science. Both instructors and students found the experience satisfactory

and did not find any difficulty in shifting from traditional assignments to our proposal.

Any peer assessment process requires that the instructor(s) must provide a rubric, i.e., clear instructions for the peer assessors to evaluate their peers' assignments. In addition to this requirement, our method also requires the instructor(s) to provide a list of aspects (labels) to be considered for deciding the overall grade. The achievement of these aspects must be assessed using either a real value (score) or a range of ordinal values. Therefore, the approach presented in this paper is applicable in the assessment of assignments in any field or discipline, provided these requirements are fulfilled.

The work presented in this paper can be extended in several aspects. Firstly, we may wish to include the grades provided by a professional instructor together with those provided by peers. The challenge will be how to merge all these grades, bearing in mind that the reliability of the instructor's grades is higher than that of the students. This extension could eventually lead us to complicate the linear model induced by our approach. We could accordingly devise a deep learning model for this purpose and check whether it improves the results. Finally, we would like to check the scalability of the models, applying them to very large datasets originating from MOOCs.

Acknowledgments

The research reported in this paper has been supported in part under grant TIN2015-65069-C2-2-R from the Ministerio de Economía y Competitividad and partially funded by European Regional Development Funds (ERDF). We would also like to acknowledge the students and instructors who collaborated with us in the assignments of our university, the University of Oviedo, in Spanish Economy (Juan Vázquez), Constitutional Law (Francisco Bastida) and Accounting Information (Mónica Álvarez Pérez).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Bahamonde, A., Bayón, G. F., Díez, J., Quevedo, J. R., Luaces, O., del Coz, J. J., Alonso, J., Goyache, F., July 2004. Feature subset selection for learning preferences: A case study. In: Greiner, R., Schuurmans, D. (Eds.), Proceedings of the International Conference on Machine Learning (ICML '04). Banff, Alberta (Canada), pp. 49–56.

Caruana, R., 1997. Multitask learning. Machine Learning 28 (1), 41–75.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1724–1734.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 (6), 391–407.

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

Formanek, M., Wenger, M. C., Buxner, S. R., Impey, C. D., Sonam, T., 2017. Insights about large-scale online peer assessment from an analysis of an astronomy MOOC. Computers & Education 113, 243 – 262.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., Struyven, K., 2010. Improving the effectiveness of peer feedback for learning. Learning and Instruction 20 (4), 304 – 315.

Guetl, C., 2007. Moving towards a fully automatic knowledge assessment tool. International Journal of Emerging Technologies in Learning (iJET) 3 (1).

Hovardas, T., Tsivitanidou, O. E., Zacharia, Z. C., 2014. Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. Computers & Education 71, 133 – 152.

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

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S. R., 2015. Peer and self assessment in massive online classes. In: Plattner, H., Meinel, C., Leifer, L. (Eds.), Design Thinking Research. Understanding Innovation. Springer International Publishing, pp. 131–168.

Labutov, I., Studer, C., 2016. JAG: Joint Assessment and Grading. In: Machine Learning for Digital Education and Assessment Systems, ICML 2016 Worshop.

Le, Q. V., Mikolov, T., 2014. Distributed Representations of Sentences and Documents. In: Proceedings of the 31st International Conference on Machine Learning. Vol. 32. pp. 1188–1196.

Liu, N.-F., Carless, D., 2006. Peer feedback: the learning element of peer assessment. Teaching in Higher education 11 (3), 279–290.

Lu, J., Law, N., 2012. Online peer assessment: Effects of cognitive and affective feedback. Instructional Science 40 (2), 257–275.

Luaces, O., Díez, J., Alonso, A., Troncoso, A., Bahamonde, A., 2015a. Including content-based methods in peer-assessment of open-response questions. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW). IEEE, pp. 273–279.

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

Luaces, O., Díez, J., Alonso-Betanzos, A., Troncoso, A., Bahamonde, A., 2017. Content-based methods in peer assessment of open-response questions to grade students as authors and as graders. Knowledge-Based Systems 117, 79–87.

Luaces, O., Díez, J., Joachims, T., Bahamonde, A., 2015c. Mapping preferences into euclidean space. Expert Systems with Applications 42 (22), 8588 – 8596.

Noorbehbahani, F., Kardan, A., 2011. The automatic assessment of free text answers using a modified bleu algorithm. Computers & Education 56 (2), 337 – 345.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 311–318.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D., 2013. 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, pp. 153–160.

Raman, K., Joachims, T., 2014. Methods for ordinal peer grading. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14. ACM Press, New York, New York, USA, pp. 1037–1046.

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

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

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

Salton, G., Wong, A., Yang, C. S., Nov. 1975. A vector space model for automatic indexing. Commun. ACM 18 (11), 613–620.

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

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

Sun, D. L., Harris, N., Walther, G., Baiocchi, M., 2015. Peer assessment enhances student learning: The results of a matched randomized crossover experiment in a college statistics class. PloS one 10 (12), e0143177.

The Theano Development Team, Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., Bengio, Y., Bergeron, A., Bergstra, J., Bisson, V., Snyder, J. B., Bouchard, N., Boulanger-Lewandowski, N., Bouthillier, X., de Brébisson, A., Breuleux, O., Carrier, P.-L., Cho, K., Chorowski, J., Christiano, P., Cooijmans, T., Côté, M.-A., Côté, M., Courville, A., Dauphin, Y. N., Delalleau, O., Demouth, J., Desjardins, G., Dieleman, S., Dinh, L., Ducoffe, M., Dumoulin, V., Kahou, S. E., Erhan, D., Fan, Z., Firat, O., Germain, M., Glorot, X., Goodfellow, I., Graham, M., Gulcehre, C., Hamel, P., Harlouchet, I., Heng, J.-P., Hidasi, B., Honari, S., Jain, A., Jean, S., Jia, K., Korobov, M., Kulkarni, V., Lamb, A., Lamblin, P., Larsen, E., Laurent, C., Lee, S., Lefrancois, S., Lemieux, S., Léonard, N., Lin, Z., Livezey, J. A., Lorenz, C., Lowin, J., Ma, Q., Manzagol, P.-A., Mastropietro, O., McGibbon, R. T., Memisevic, R., van Merriënboer, B., Michalski, V., Mirza, M., Orlandi, A., Pal, C., Pascanu, R., Pezeshki, M., Raffel, C., Renshaw, D., Rocklin, M., Romero, A., Roth, M., Sadowski, P., Salvatier, J., Savard, F., Schlüter, J., Schulman, J., Schwartz, G., Serban, I. V., Serdyuk, D., Shabanian, S., Simon, É., Spieckermann, S., Subramanyam, S. R., Sygnowski, J., Tanguay, J., van Tulder, G., Turian, J., Urban, S., Vincent, P., Visin, F., de Vries, H., Warde-Farley, D., Webb, D. J., Willson, M., Xu, K., Xue, L., Yao, L., Zhang, S., Zhang, Y., may 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.0.

Thomas, P., Haley, D., DeRoeck, A., Petre, M., 2004. E-assessment using latent semantic analysis in the computer science domain: A pilot study. In: Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning. Association for Computational Linguistics, pp. 38–44.

Tseng, S.-C., Tsai, C.-C., 2007. On-line peer assessment and the role of the peer feedback: A study of high school computer course. Computers & Education 49 (4), 1161 – 1174.

Wang, H.-C., Chang, C.-Y., Li, T.-Y., 2008. Assessing creative problem-solving with automated text grading. Computers & Education 51 (4), 1450 – 1466.

Appendix A. Derivatives used in the gradient descent

The learning method proposed in this paper is based on Stochastic Gradient Descent (SGD). The idea is to find the parameters of the model being learned which minimize a given loss function. This is an iterative method which proceeds by presenting randomly selected training examples, evaluating the output (comparing the prediction with the ground truth) and modifying the parameters of the model in the direction that minimizes the error (loss function).

In each iteration, the parameters of the model, Θ (in our case the matrices $\boldsymbol{W}$ and $\boldsymbol{V}$) are modified, as follows:

$$\Theta \leftarrow \Theta - \gamma \left(\frac{\partial \operatorname{err}(\Theta)}{\partial \Theta} + \nu \cdot \frac{\partial \|\Theta\|_F^2}{\partial \Theta} \right), \tag{A.1}$$

where $\|\cdot\|_F^2$ is the Frobenius norm included for *regularization*, γ is the *learning rate* and ν is the *regularization factor*. As usual, γ decreases its value in each iteration. In the experiments reported in Section 4 we have used the following expression to determine the value of γ in the i^{th} iteration:

$$\gamma = \frac{1}{1 + \gamma_s \cdot i}. \tag{A.2}$$

The derivatives used in the gradient descent, when the maximum in the loss function (8) is greater than zero, are given by

$$\frac{\partial \operatorname{err}(\Theta)}{\partial \boldsymbol{W}} = \boldsymbol{V}(\boldsymbol{a}^w - \boldsymbol{a}^b)((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l})^{\mathrm{T}} \tag{A.3}$$

$$\frac{\partial \operatorname{err}(\Theta)}{\partial \boldsymbol{V}} = \boldsymbol{W}((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l})(\boldsymbol{a}^w - \boldsymbol{a}^b)^{\mathrm{T}} \tag{A.4}$$

and the derivative used in the regularization term is

$$\frac{\partial \|\boldsymbol{X}\|_F^2}{\partial \boldsymbol{X}} = 2\boldsymbol{X} \tag{A.5}$$

where $\boldsymbol{X}$ is either $\boldsymbol{W}$ or $\boldsymbol{V}$.

The derivatives (A.3), (A.4) and (A.5) included in this appendix are useful for those who want to implement their own version of gradient descent, but it is worth noting that modern machine learning frameworks, like Theano (The Theano Development Team et al., 2016) or TensorFlow (Abadi et al., 2016), include automatic differentiation, as well as several improved versions of optimization algorithms, so the developer does not need to know or program these derivatives.