Exploiting Collaborative Filtering Techniques for Automatic Assessment of Student Free-text Responses

Tao Ge
Key Laboratory of Computational Linguistics, Ministry of Education, School of Electronics Engineering and Computer Science, Peking University, China
getao@pku.edu.cn

Zhifang Sui*
Key Laboratory of Computational Linguistics, Ministry of Education, School of Electronics Engineering and Computer Science, Peking University, China
szf@pku.edu.cn

Baobao Chang
Key Laboratory of Computational Linguistics, Ministry of Education, School of Electronics Engineering and Computer Science, Peking University, China
chbb@pku.edu.cn

ABSTRACT

The automatic assessment of free-text responses of students is a relatively newer task in both computational linguistics and educational technology. The goal of the task is to produce an assessment of student answers to explanation and definition questions typically asked in problems seen in practice exercises or tests. Unlike some conventional methods which assess the student responses based on only information about their corresponding questions, this paper exploits idea of collaborative filtering to analyze student responses and used an effective collaborative filtering model – feature-based matrix factorization model to deal with this challenge. The experimental results show that our feature-based matrix factorization model outperforms the baseline models and the model with a re-ranking phase can achieve a better and competitive performance – 63.6% overall accuracy on the Beetle dataset.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

General Terms

Algorithms, Experimentation

Keywords

Assessment of student response, collaborative filtering, feature-based matrix factorization

1. INTRODUCTION

As the Internet technology develops at a staggering rate, an increasing number of tests such as TOEFL-IBT are taken online instead of in the traditional paper-based pattern. Compared with paper-based test (PBT), computer-based test (CBT) and Internet-based test (IBT) can save much labor of human raters and resources. Although computers are capable of assessing answers of students to some types of questions like multiple choice questions, human raters are still indispensable at present because it is a challenge for computers to accurately assess free-text responses of students to some questions such as English writing tasks. Also, current e-learning systems have limited capability for giving students feedback and providing automatic assessment since there is no established technology for assessing natural language responses to questions. Therefore, automatic assessment of student answers is worthy of investigation.

The task of automatic assessment of student responses proposed in semeval-2013 [1] tries to deal with the challenge of automatic assessment of student responses. The goal of the task is to grade student answers for enabling well-targeted and flexible feedback in a tutorial dialogue setting.

Previous work on student answer assessment for intelligent tutoring systems used LSA [2], classifiers based on "bag-of-words" features [3] to determine if a student answer corresponds to one of the expected correct or incorrect answers anticipated by a system designer. More recently, [4, 5] formulated the problem of assessing student input in terms of recognizing textual entailment.

However, previous methods for this task predict the grading level of a given response based on only information about its corresponding question such as reference answers and grading level of responses to this question. In other words, they do not take into consideration the grading information about responses to other questions. Unlike these conventional approaches, this paper exploits the idea of collaborative filtering to analyze the student responses, which predicts the grading level of a student response based on both the grading records of its corresponding question and the global information of gradings across the dataset. It is not difficult to understand the fact that the global grading information is useful for accurately assessing the responses to a specific question because this information can tell us which grading level the most responses get and what kind of responses tend to get high or poor grades. For instance, a response whose text is “I don’t know” is always a poor response to whatever questions and this fact can help accurately predict such responses to a unseen question in the test set. Furthermore, in a recommendation perspective, if two users have similar shopping records, then they may have the similar preferences for items; likewise, some questions may have similar “preference” for

*This author is the corresponding author.

CIKM’13, October 27 – November 01 2013, San Francisco, CA, USA

ACM 978-1-4503-2263-8/13/10$15.00.
http://dx.doi.org/10.1145/2505515.2507827 .

semantic information of student responses. Assuming that good responses to given two questions always share many features, if a response to one of the questions is similar to a good response to the other question, then it is likely to be a good response to its corresponding question. Based on this intuition, we consider this task as a rating prediction problem where we try to predict the "rating" of questions (users) to student responses (items) by using a popular collaborative filtering model – feature-based matrix factorization model.

2. TASK DESCRIPTION

Since it is a relatively newer task in both knowledge management and educational technology community, we briefly describe the task of automatic assessment of student responses. The goal of this task is to assess student answers to exercise questions that can be useful in tutorial dialogue and/or e-learning systems. Specifically, given a question, a known correct "reference answer", a set of student answers with manually annotated grading levels and a 1- or 2-sentence student answer, the goal is to determine the student's answer accuracy.

The task can be performed at different levels of granularity. In this paper, we mainly address the 5-way task, where the system is required to classify the student answer according to one of the following judgments:

- $Correct$, if the student answer is a complete and correct paraphrase of the reference answer;
- $Partially\_correct\_incomplete$, if the student answer is a partially correct answer containing some but not all information from the reference answer;
- $Contradictory$, if the student answer explicitly contradicts the reference answer;
- $Irrelevant$, if the student answer is "irrelevant", talking about domain content but not providing the necessary information;
- Non_domain, if the student answer expresses a request for help, frustration or lack of domain knowledge - e.g., "I don't know", "as the book says", "you are stupid".

In this paper, we used the Beetle dataset [6] for training and evaluation, which is a set of transcripts of students interacting with an intelligent tutorial dialogue system for teaching conceptual knowledge in the basic electricity and electronics domain.

3. COLLABORATIVE FILTERING MODEL FOR ASSESSING RESPONSES

In this section, we discuss how to use collaborative filtering techniques to deal with the challenge. We first explain why collaborative filtering techniques can be exploited for this task in Section 3.1. Then, a popular collaborative filtering model – feature-based matrix factorization model is to be discussed in detail in Section 3.2. Finally, Section 3.3 presents a re-ranking method for re-ranking the marginal predictions.

3.1 Motivation

Collaborative filtering is one of the most promising technologies for recommender systems. In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly. This intuition is very popular in the online shopping recommendation.

Likewise, the task of automatic assessment of student responses can also be addressed using the idea of collaborative filtering. For a given question, different student responses may be graded with different grading levels just as different items might be rated with different ratings by a given user under the online shopping scenario.

For a specific example, there are two questions whose ids are q1 and q2 respectively. Question of q1 is "why are wires made of copper?" and question of q2 is "Can silver conduct? If so, why can we hardly see wires made of silver?". By analyzing the student responses, we can find responses involving the issue of conductivity and price to both q1 and q2 are graded with a high grade such as $Correct$. In contrast, responses which do not involve these two aspects are graded with a poor grade like $Irrelevant$ for the two questions. Based on the facts mentioned above, assuming that r1 and r2 are a response to q1 and q2 respectively and they are very similar (e.g. similar unigram features), if r1 is a good response to q1, then a response r2 which is similar to r1 is very likely to be a good response to q2 and vice-versa because q1 and q2 have the similar "preferences" for the semantic information concerning "conductivity" and "price", which is exactly the idea of collaborative filtering.

Also, there are some responses such as "I don't know" which are always graded with a poor grade for whatever questions. For a collaborative filtering model, a negative bias will be assigned to such responses. As a result, even for a new question without any prior information or reference answers, such responses will be graded with a poor grading level by the collaborative filtering model, which can hardly be handled by other previous models.

For the above-mentioned reasons, we used a typical and effective collaborative filtering model – feature-based matrix factorization model to address the challenge. The feature-based matrix factorization model can predict the grade of a response based on both the grading records of its corresponding question and the global information of gradings across the dataset. Furthermore, this model can capture semantic information with a latent factor space, which also contributes to a good performance.

3.2 Feature-based Matrix Factorization Model for Automatic Assessment of Responses

One of collaborative filtering models is matrix factorization models [7]. Matrix factorization models transform both items and users to the same latent factor space which tries to explain ratings by characterizing both products and users on factors automatically inferred from user feedback, as shown in Figure 1.

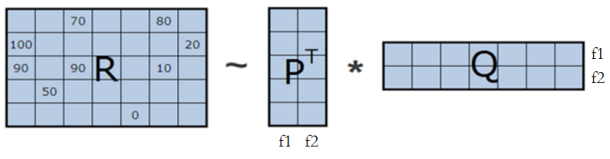

Figure 1: Intuition of Matrix Factorization Models

Previous research has shown that the matrix factorization model can model recommendation problems well. However, it is not difficult to find that the task of predicting rating a user rates to an item is very similar to the task of assessing student responses given a question.

We consider the assessing task as a rating prediction problem. Specifically, we regard questions as users, responses as well as reference answers as items and grading levels as ratings. For the 5-way task, the grading levels *Correct*, *Partially_correct_incomplete*, *Contradictory*, *Irrelevant* and *Non_domain* are mapped to the rating space $\mathbf{R}$ which is defined as follows:

$$\mathbf{R} = \{1, 2, 3, 4, 5\}$$

In our task, each row of the rating matrix R in Figure 1 represents a question, each column represents a response and each element of the matrix is the grade level of a response(column) to a question(row). For the factor matrice P and Q, the dimension f_i is a latent semantic dimension. For the example mentioned by Section 3.1, f_1 might represent the semantic information about conductivity and f_2 might represent the information about price. If the value of f_1 of a question is large, then it means that the desirable answer to this question should contain sufficient information about conductivity. On the other hand, if the value of f_1 of a response is large, then it means that the response provides much information about conductivity. In this way, some complex information can be captured by the latent factor spaces and also "preference" of a question for a response is modeled.

However, basic matrix factorization models such as basic SVD cannot address this challenge well because these models do not use any features about questions and responses except their IDs. Since each response corresponds to only one question, features are extremely sparse and responses to different questions do not have any connection even if their text is very similar. For example, assuming that there are two responses to two different questions and text of these two responses is the same, e.g. their text is "I don't know", but the model cannot be aware of that the responses are the similar because their features i.e. their IDs are totally different and they are never "rated" by the same questions. Likewise, similar questions also cannot be identified by this model. As a result, the basic matrix factorization model cannot work at all.

For solving the problem mentioned above, we have to introduce more features for profiling questions and answers in order to make features less sparse. For leveraging more features of questions and responses, feature-based matrix factorization model which was proposed by [8] is used. This model generalizes the basic factorization models, in which new types of information can be utilized by simply defining new features. The framework of feature-based matrix factorization is shown by Equation 1 where μ is a constant indicating global mean value of rating, $b^{(g)}$, $b^{(u)}$ and $b^{(i)}$ are biases of global features, user features and item features respectively, $\boldsymbol{p}$ and $\boldsymbol{q}$ represent factors of features of users and items respectively and α, β and γ are weights of user features, item features and global features respectively.

$$\hat{r} = \mu + \sum_j b_j^{(g)}\gamma_j + \sum_j b_j^{(u)}\alpha_j + \sum_j b_j^{(i)}\beta_j + (\sum_j \boldsymbol{p_j}\alpha_j)^T \cdot (\sum_j \boldsymbol{q_j}\beta_j) \quad (1)$$

To model profiles of questions and responses with features, we select $questionid$ as the feature of questions and bag-of-words of responses as the response features. The reason why we do not select text of questions as features is that the question text is very confusing. For example, text of many questions is a word such as "why". The specific feature-based matrix factorization model for our task is shown in Equation 2, in which $S(i)$ is the set of features of the response i.

$$\hat{r}_{u,i} = \mu + b_u + \sum_{j \in S(i)} b_j\beta_j + \boldsymbol{p_u}^T \cdot (\sum_{j \in S(i)} \boldsymbol{q_j}\beta_j) \quad (2)$$

3.3 Re-ranking the marginal predictions

Feature-based matrix factorization is naturally a regression model so its prediction of each test example is a numeric instead of a nominal class. Although we can use rounding-off method to remap the numeric to class label, it is not effective for some cases. For a test example predicted with a marginal score e.g. 4.5, it is difficult to tell whether it is should be graded as *Correct* or *Partially_correct* for the matrix factorization model. Therefore, we used a maximum entropy classifier to re-rank such marginal predictions for a better performance.

We select bag-of-word features and train the MaxEnt classifier. For the test example $t \in T$, MaxEnt classifier serves re-predicting their classes. T is defined as follows and $pred(t)$ is the predicting score by the matrix factorization model.

$$T = \left\{ t \middle| |pred(t) - \lfloor pred(t) + 0.5 \rfloor| > 0.3 \right\}$$

4. EXPERIMENTS AND EVALUATIONS

In this section, we first introduce the experimental settings in detail. Then we discuss the results and give an analysis.

4.1 Experimental Setting

Dataset The dataset we used for evaluation is the Beetle Dataset, which contains 47 questions. Each question is associated with 1 to 10 different reference answers provided by experienced tutors and dozens of student responses.

Pre-processing In the pre-processing step, we perform lemmatization for each token and filter out stop words.

Evaluation Since the semeval-2013 organizer has not released the test set with golden standard, we alternatively perform cross validation to evaluate the performance of our model. To simulate the scenario of unseen answers [1], we randomly divided the student responses into 20 groups and perform 20-fold cross validation. The test set in each fold contains 197 or 198 test examples of which there are on average 4.2 student responses to each question.

4.2 Experimental Results

We set the following models which only use lexical features as baselines for evaluations:

Baseline1: Majority Class The majority class baseline is a model which assigns $Correct$ (the most frequent class) to each test instance.

Baseline2: Lexical Similarity The lexical similarity baseline is a simple decision tree classifier with features such as lexical similarity and overlap by using an implementation toolkit – Weka [9].

Baseline3: MaxEnt Classifier This baseline is a maximum entropy classifier using bag-of-word features. The classifier also serves re-ranking the marginal predictions, as discussed in Section 3.3.

Note that Baseline1 and Baseline2 are baselines officially used in the semeval-2013 competition. We compare the performance of following models with the baselines:

Model1: Feature-based matrix factorization model which predicts grading levels of student responses. We used the rounding-off method to map the numeric prediction to one of the five given classes.

Model2: Feature-based matrix factorization model with re-ranking. The classifier for re-ranking is the model described by Baseline3.

Table 1 shows the performance of different models for the task of student response analysis. For saving space, we use integers to represent the grade levels in Table 1 in which "macro" and "micro" mean macro-average and micro-average respectively and "overall" represents the overall accuracy of models.

The majority baseline achieves 42.3% overall accuracy. However, this is obviously at the expense of serious errors. For instance,

Grade	B1			B2			B3			M1			M2		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
5	0.423	**1.000**	0.594	0.645	0.778	0.705	0.667	0.802	0.728	**0.808**	0.735	0.769	0.754	0.787	**0.770**
4	0.000	0.000	0.000	0.415	0.345	0.377	**0.485**	0.454	0.469	0.401	**0.610**	0.484	0.460	0.554	**0.503**
3	0.000	0.000	0.000	0.440	0.364	0.398	0.602	0.507	0.551	**0.652**	0.484	0.556	0.629	**0.524**	**0.572**
2	0.000	0.000	0.000	0.093	0.035	0.051	**0.286**	**0.248**	**0.265**	0.123	0.150	0.135	0.234	0.230	0.232
1	0.000	0.000	0.000	0.640	**0.846**	**0.728**	0.959	0.482	0.642	0.946	0.544	0.691	**0.982**	0.569	0.721
macro	0.084	0.200	0.119	0.447	0.474	0.452	0.600	0.499	0.531	0.586	0.505	0.527	**0.612**	**0.533**	**0.560**
micro	0.178	0.422	0.251	0.521	0.549	0.529	0.611	0.611	0.603	**0.659**	0.613	0.624	0.649	**0.636**	**0.637**
overall		0.423			0.549			0.611			0.613			**0.636**	

Table 1: Performance of different models for the task of student response analysis

such a system would tell the students that they are correct even if they are saying something contradictory. This is reflected in a much lower macro-averaged F score. Compared with the majority baseline, Baseline2 and Baseline3 can assess student responses more accurately and achieve the overall accuracy 54.9% and 61.1% respectively but their abilities to capture semantic of responses seem not so good as the feature-based matrix factorization model which tries to represent the semantic information in a latent factor space and achieved a performance of 61.3% overall accuracy.

Furthermore, it is clear that combining a re-ranking phase can improve the performance of feature-based matrix factorization models for the reason that errors due to marginal predictions made by the matrix factorization model are corrected. The matrix factorization model with re-ranking can achieve 63.6% overall accuracy, which is a competitive performance for a model which only exploits the bag-of-word features without too much pre-processing such as spelling correction for the task.

It is also notable that responses which are *Correct* or *Non_domain* can be identified most easily by all models except the majority baseline since the responses of these two grades have more distinct features than ones of other grades. In contrast, it seems quite difficult for models to identify irrelevant responses since such responses may contain some important words mentioned in either reference answers or good responses. As a result, they are likely to be graded with a high grade. Therefore, identifying such responses requires deeper semantic analysis. In addition, it can be found that the matrix factorization model is very awkward in identifying the irrelevant responses though it is good at dealing with other levels of responses. In contrast, the maximum entropy classifier has a more stable performance, which is reflected in a higher macro-average F score. Therefore, when the matrix factorization model is combined with the MaxEnt classifier, its weakness in handling the irrelevant responses can be addressed to some extent and that is probably one of reasons why combining a MaxEnt classifier helps improve the performance of the matrix factorization model.

5. CONCLUSIONS

This paper addresses the task of automatic assessment of student responses in a novel perspective. We model the problem as a rating prediction problem and used a feature-based matrix factorization model to predict the grading levels of responses to their corresponding questions based on the idea of collaborative filtering. The experiments show that the feature-based matrix factorization model can deal with the assessment task well. Furthermore, the feature-based matrix factorization model combined with a re-ranking phase achieves a higher performance – 63.6% overall accuracy – a competitive performance for a model only using bag-of-word features for this task, which proves the effectiveness of our model. Additionally, our model is so flexible that we can also exploit more promising features such as n-gram features, categories of questions, syntactic features in this model for better performance, which is to be explored in future work.

Acknowledgements

This paper is supported by NSFC Project 61075067 and National Key Technology R&D Program (No: 2011BAH10B04-03).

6. REFERENCES

[1] Myroslava O Dzikovska, Rodney D Nielsen, and Chris Brew. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210. Association for Computational Linguistics, 2012.

[2] Art Graesser, Phanni Penumatsa, Matthew Ventura, Zhiqiang Cai, and Xiangen Hu. Using lsa in autotutor: Learning through mixed-initiative dialogue in natural language. 2007.

[3] Pamela W Jordan, Maxim Makatchev, and Kurt VanLehn. Combining competing language understanding approaches in an intelligent tutoring system. In *Intelligent Tutoring Systems*, pages 346–357. Springer, 2004.

[4] R Nielsen, Wayne Ward, James H Martin, and Martha Palmer. Annotating students' understanding of science concepts. In *Proc. LREC*, 2008.

[5] Rodney D Nielsen, Wayne Ward, and James H Martin. Learning to assess low-level conceptual understanding. In *Proceedings 21st International FLAIRS Conference, Coconut Grove, Florida, May*, 2008.

[6] Myroslava O Dzikovska, Diana Bental, Johanna D Moore, Natalie B Steinhauser, Gwendolyn E Campbell, Elaine Farrow, and Charles B Callaway. Intelligent tutoring with natural language support in the beetle ii system. In *Sustaining TEL: From Innovation to Learning and Practice*, pages 620–625. Springer, 2010.

[7] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[8] Tianqi Chen, Zhao Zheng, Qiuxia Lu, Weinan Zhang, and Yong Yu. Feature-based matrix factorization. *arXiv preprint arXiv:1109.2271*, 2011.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.