Computer-Aided Response-to-Intervention for Reading Comprehension Based on Recommender System

Ming-Chi Liu[1], Wei-Yang Lin[1], and Chia-Ling Tsai[2(✉)]

[1] Chung Cheng University, Chiayi, Taiwan
[2] Queens College, CUNY, Queens, NY 11367, USA
ctsai@qc.cuny.edu

Abstract. In 2019, New York State Education Department announced 54.6% of all students in grades 3 to 8 not meeting the standard of reading proficiency. Motivated by the need for a more efficient intervention model, we propose a recommender system to leverage the technology in machine learning to recommend suitable reading materials for effective intervention. The recommendation is based on the student's prior reading comprehension assessments and also assessments of other students at the same grade level using collaborative filtering. No other prior academic or demographic information of students is available. Two main challenges are lack of explicit ratings of reading passages by students and the small data size. Both are addressed in this paper. BERT is applied to determine the textual evidence of a question, and linguistic properties are extracted to generate a continuous rating for a question answered by a student to reflect the skill level of the student. The difficulty level of a passage is determined by the associated multiple-choice questions. The system is trained with a collection of fourth grade New York English Language Arts assessments. The training dataset is augmented with synthetic data using SMOTE for better generalizability. Our system achieves 75.7% in accuracy and 59.23% in F1-score.

Keywords: Reading comprehension · Intervention program · Recommender system · Collaborative filtering

1 Introduction

In New York State, USA, students in grades 3 to 8 take the State English Language Arts (ELA) test each spring. An ELA test contains multiple-choice questions and open-ended questions based on short passages in the test. To do well, students should be able to read the text closely for textual evidence and to make logical inferences from it. In 2019, New York State Education Department reported that 54.6% of all students in grades 3 to 8 do not meet the standard

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13355, pp. 512–518, 2022.
https://doi.org/10.1007/978-3-031-11644-5_42

of proficiency [2]. To drive the changes in students who are at some level of risk for not meeting academic expectation, schools arrange academic intervention service for all students who are well below or partially proficient, based on the ELA score. However, a single performance score is not instrumental in explaining the lack of specific language knowledge and skills typically demonstrated at that grade level.

Our work is motivated by the need for a more effective Response-to-Intervention model that continuously assess the need for changes in instruction and goals, driven by students' progress data [6]. We develop a machine learning (ML) system for recommending instruction materials for reading comprehension, based on the prior reading assessments of an individual student and also of students in the same grade level. A recommender system for e-learning comes in various formats, depending on the data availability. A common approach is to make recommendations of courses or predictions of student performance based on known student and course characteristics [7–9]. Such problem can be easily formulated as a classification or regression problem. Thai-Nghe et al. [10] proposed a recommender system for math assessment based on matrix factorization where student factors and some of the problem factors are known. Our work is most similar to [10], but performs recommendation of reading passages based on the predicted ratings of associated multiple-choice questions by a given student, without any prior information regarding either the student or the assessment material.

There are two main challenges to be addressed. First, the multiple-choice questions have only dichotomous ratings (correctly or incorrectly answered) from students. Second, the performance of a recommender system is limited by the prior data for training, but our dataset is relatively small, comparing to the growing dataset of millions of records for a commercial recommender system. We explore linguistic properties of reading passages to address the first challenge and employ data augmentation for the second challenge.

2 Dataset

The dataset consists of two parts. The first part is the set of six fourth grade New York State mock ELA examinations from year 2005 to 2010. Only multiple-choice questions are considered. In each examination there are five passages, each having five to six associated questions, for a total of 28 questions. Question and choice statements are typically short. The second part is the set of student assessments, involving a total of 378 randomly-selected fourth grade students with various levels of reading proficiency from 17 reading intervention classes. Every participant was assigned an identification number and participated in most three mock examinations. Each student answered from 4 to 84 questions in total. No other background information, such as prior academic performance or demographic data, was collected to protect the privacy of the participants.

3 Methodology

A recommender system can be understood as a ML problem trained with a set of 3-tuples: {[user ID, product ID, rating]}. Our input is a tuple of [student ID, question ID, student answer]. As shown in Fig. 1, the weights of features associated with the j-th student and with the i-th question should be estimated jointly by the recommender system, and the rating of a new student-question pair can be predicted using the estimated features. The fitness level of a reading passage can be computed as the average of the predicted ratings of associated multiple-choice questions by a given student. The system makes recommendation by choosing passages with the rating in the range $[0 - \epsilon, 0]$—a difficulty level slightly surpassing the recent strength of a student—to promote learning.

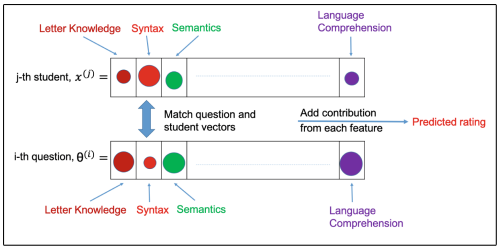

Fig. 1. General concept of a recommender system demonstrated for e-learning. Student features and question features are estimated from known ratings, and a new rating is predicted as a combination of the estimated features.

To generate the ratings for student-question pairs, neither the nominal answer choices nor the dichotomous scoring outcome of 0/1 is appropriate for a recommendation system. It is necessary to convert the student's answer choice of a question to a continuous rating, reflecting the skill level of a student for answering a given question. Data augmentation is applied to the dataset to increase the data variation for better generalizability of the model. Collaborative filtering is adopted as the filtering technique of the recommendation system, which simultaneously computes the student features and the question features using all available ratings in the training pool.

3.1 Rating Transformation

To transform the rating from either nominal or dichotomous to continuous for better discrimination, we explore ML techniques in natural language processing to connect the questions to the passage with the textual evidence, and to extract linguistic properties of the questions. A rating computed from the linguistic properties of a question should reflect the underlying language skills needed to correctly answer the given question.

To locate the textual evidence of a question in the passage, Bidirectional Encoder Representation from Transformer (BERT) [5] is applied to identify connection between a question and a sentence in the passage. The pre-trained version used in this study is BERT-base-uncased, which is further fine-tuned with OneStopQA dataset [3] for reading comprehension with multiple-choice questions.

Combining with the textual evidence, linguistic features are identified for each answer choice of a question, because linguistic properties have been shown to be important indicators for readability of a passage. We explore the Suit for Automatic Linguistic Analysis Tools (SALAT) [1] for the social sciences. There is a total of 3908 features generated for each answer choice (combined with the textual evidence) of a question. Values of a feature are normalized to Z-scores using the mean and standard deviation of the feature.

To generate the rating of a student-question pair, the 3908-tuple feature vector of the student's answer choice is converted into a scalar value using the magnitude of the vector, which is normalized to a Z-score again and scaled to the range of $[0, 1]$, using a sigmoid function for linear scaling mainly in the clustered sections closer to 0. Let d_f be the magnitude of the feature vector f in Z-score, the sigmoid score $s(d_f) \in [0, 1]$ is computed as $s(d_f) = \frac{1}{1+\exp(c*(-d_f))}$, where $c = 2.5$ determined empirically. The final rating $r = s(d_f)$ is set for a correct answer choice and $r = s(d_f) - 1$ for an incorrect answer choice, so all correct answers have ratings greater than 0 and all incorrect answers have ratings less than 0. $r \in [-1, 1]$ with 1 representing exceptional and –1 for well below proficient.

3.2 Data Augmentation

Data augmentation is common in ML to increase the amount of data to deal with the problem of class imbalance or to improve generalizability of the model. To add synthetic data that are slightly modified of existing data, we adopt SMOTE (Synthetic Minority Oversampling Technique) [4] to generate additional samples for each student. Given a question that a student completed, the feature vector f is computed as the average of the feature vectors of the 4 answer choices. Based on the cosine similarity, its K (=3) nearest neighbors of the same student are located, and one neighbor f' is randomly decided to determine the direction of perturbation. The new feature $\bar{f}$ is f perturbed with a random portion η between [0,1] of the difference between f and f': $\bar{f} = f + \eta(f' - f)$. $\bar{f}$ is marked as correctly or incorrectly answered question the same as f, and is converted to the rating following the same steps described in Sect. 3.1.

3.3 Collaborative Filtering

Collaborative filtering implemented using matrix factorization aims to estimate two types of feature vectors: $x^{(i)} \in \Re^n$ representing i-th question and $\theta^{(j)} \in \Re^n$ representing j-th student, where $n = 10$ set empirically. There are n_m questions and n_u students. $x^{(i)}$ is constrained by ratings of all students who answered

question i and $\theta^{(j)}$ is constrained by ratings of all questions that student j answered. When both $[x^{(1)}, \ldots, x^{(n_m)}]$ and $[\theta^{(1)}, \ldots, \theta^{(n_u)}]$ are unknown, they can be estimated jointly by minimizing the following cost function:

$$J\left(x^{(1)}, .., x^{(n_m)}, \theta^{(1)}, .., \theta^{(n_u)}\right) = \frac{1}{2} \sum_{(i,j):r(i,j)=1} \left(\left(\theta^{(j)}\right)^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^{n} \left(x_k^{(i)}\right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^{n} \left(\theta_k^{(j)}\right)^2, \tag{1}$$

where $y^{(i,j)}$ is the rating for student j and question i, and $r(i,j) = 1$ indicates valid rating for student j and question i. The last two terms in Eq. 1 are for regularization to avoid overfitting. The minimization process is initialized with random values for all vectors and alternates the estimations of $[x^{(1)}, \ldots, x^{(n_m)}]$ and $[\theta^{(1)}, \ldots, \theta^{(n_u)}]$ by fixing another vector until the process converges.

To predict whether student j will correctly answer question i, the rating r is estimated as $r = (\theta^{(j)})^T x^{(i)}$. To map r back to the answer choices, ratings of the choices are computed and the choice with the rating closest to r is the predicted choice of student j if given question i.

4 Experiments and Results

There is a total of 27731 answered records with 67.5% correctly answered. Data augmentation was applied to add another 35773 records for training—168 questions per student in total. Leave-one-out validation was performed on only the original dataset of 27731 records. As a limitation of our current study, we were able to only validate the performance prediction of a student on a question, not the effectiveness of the recommendation for Response-to-Intervention.

We assessed the performance of the system with accuracy and F1-score. Since incorrectly answered records are considered the minority for the classification problem, they are considered the positive class, whereas the correctly answered records are the negative class. A prediction is considered correct if the student scored or not scored the question and system predicts the same outcome, regardless of the answer choice picked. The F1-score provides better insight to a problem with imbalance classes since it ignores the correct predictions of the majority class (i.e. true negatives) by considering only the precision and recall of the minority class. Our system achieves an F1-score of 55.93% and accuracy of 72.83% without data augmentation, and 59.23% and 75.7% with data augmentation. We also compared our system with content-based filtering as the engine of the recommender system [11]—the rating given by a student for a new question is computed from the ratings of K (=3) nearest questions answered by the same student. Collaborative filtering outperforms content-based filtering by 2.45% in accuracy and 6.34% in F1-score. It shows the importance of automatic determination of feature representation of questions from the data using all ratings available by the cohort.

If the process of rating transformation is completely removed and the continuous rating is replaced with 0/1 rating, i.e. 1 for a correctly answered question and 0 for an incorrectly answered question, the F1-score degrades substantially from 55.93% to 27.78%. The use of BERT for evidence identification only improves the F1-score by 1.59%. An explanation for the very minor improvement from BERT is lack of discrimination of strong evidence supporting a question; on average, close to 54% of a passage is considered as the evidence for an answer choice of a question, but most statements are irrelevant. As a result, 3908-tuple feature vectors for 4 answer choices of a question can be very close in the feature space and the transformed ratings are less dispersed.

5 Conclusions and Discussion

The proposed system supports intervention for reading comprehension by recommending reading passages of difficulty level slightly surpassing the recent strength of a student to promote learning. Our proposed rating transformation scheme doubles the F1-score by converting the binary score to a continuous value using linguistic properties of a question with its supporting evidence from the reading passage. Data augmentation further boosts the performance by 3.3%. Our model can be easily generalized for other formats of reading comprehension, such as short-answer questions, if linguistic properties can be reliably computed from the associated question(s) with the supporting textual evidence identified.

Acknowledgments. This work was partially supported by grants from NSF-CHS Award 1543639, Taiwan MOST Award 109-2221-E-194-040 and PSC-CUNY Research Award 65406-00-53.

References

1. Suit for Automatic Linguistic Analysis Tools. https://www.linguisticanalysistools.org/. Accessed Dec 2021
2. New York SED: 3–8 assessment database (2019). http://www.nysed.gov/news/2019/state-education-department-releases-spring-2019-grades-3-8-ela-math-assessment-resuls. Accessed Dec 2021
3. Berzak, Y., Malmaud, J., Levy, R.: STARC: structured annotations for reading comprehension. In: Proceedings of 58th Annual Meeting of the Association for Computational Linguistics, pp. A:567–576 (2020)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv abs/1810.04805 (2019)
6. Fuchs, D., Fuchs, L.: Responsiveness-to-intervention: a blueprint for practitioners, policymakers, and parents. Teach. Except. Child. **38**(1), 57–61 (2001)
7. Goga, M., Kuyoro, S., Goga, N.: A recommender for improving the student academic performance. Procedia Soc. Behav. Scoi. **180**, 1481–1488 (2015)

8. Kurniadi, D., Abdurachman, E., Warnars, H., Suparta, W.: A proposed framework in an intelligent recommender system for the college student. J. Phys. Conf. Ser. **1402**(6), 066100 (2019). https://doi.org/10.1088/1742-6596/1402/6/066100
9. Sweeney, M., Rangwala, H., Lester, J., Johri, A.: Next-term student performance prediction: a recommender systems approach. J. Educ. Data Mining **8**(1), 22–50 (2016)
10. Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., Schmidt-Thieme, L.: Recommender system for predicting student performance. Procedia Comput. Sci. **1**(2), 2811–2819 (2010)
11. Thorat, P.B., Goudar, R.M., Barve, S.S.: Survey on collaborative filtering, content-based filtering and hybrid recommendation system. Int. J. Comput. Appl. **110**(4), 31–36 (2015)