Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning

Kang Xue[1] **Victoria Yaneva**[2] **Christopher Runyon**[2] **Peter Baldwin**[2]

[1]Department of Educational Psychology, University of Georgia, Athens, USA
KangXue@uga.edu
[2]National Board of Medical Examiners, Philadelphia, USA
{vyaneva, crunyon, pbaldwin}@nbme.org

Abstract

This paper reports on whether transfer learning can improve the prediction of the difficulty and response time parameters for $\approx$ 18,000 multiple-choice questions from a high-stakes medical exam. The type of the signal that best predicts difficulty and response time is also explored, both in terms of representation abstraction and item component used as input (e.g., whole item, answer options only, etc.). The results indicate that, for our sample, transfer learning can improve the prediction of item difficulty when response time is used as an auxiliary task but not the other way around. In addition, difficulty was best predicted using signal from the item stem (the description of the clinical case), while all parts of the item were important for predicting the response time.

1 Introduction

The questions on standardized exams need to meet certain criteria for the exam to be considered fair and valid. For example, it is often desirable to collect measurement information across a range of examinee proficiencies but this requires that question difficulties span a similar range. Another consideration is the time required to answer each question: allocating too little time makes the exam speeded whereas allocating too much time makes it inefficient. Typically, difficulty and response time measures are needed before new questions can be used for scoring. Currently, these measures are obtained by presenting new questions alongside scored items on real exams; however, this process is time consuming and costly. To address this challenge, there is an emerging interest in predicting item parameters based on item text (Section 2). The goal of this application is to filter out items that should not be embedded in live exams—even as unscored items—because of their low probability of having the desired characteristics.

In practice, there may be situations where data are available for one item parameter but not for another. For example, when a pen-and-paper test is being migrated to a computer-based test, response time measures to individual questions will not be among the historical pen-and-paper data whereas item difficulty measures will be. In this scenario, the only available response-time data would be those collected from the small sample of examinees who first piloted the computer-based test. Yet, since item characteristics like response time and difficulty are often related (e.g., more difficult items may require longer to solve), it is conceivable that information stored while learning to predict one parameter then could be used to improve the prediction of another. In this paper, we explore whether approaches from the field of transfer learning may be useful for improving item parameter modeling.

We hypothesize that transfer learning (TL) can improve the prediction of difficulty and response time parameters for a set of $\approx$18,000 multiple-choice questions (MCQs) from the United States Medical Licensing Examination (USMLE®). We present two sets of experiments, where learning to predict one parameter is used as an auxiliary task for the prediction of the other and vice versa. In addition to our interest in parameter modeling, we investigate the type of signal that best predicts difficulty and response time, which is done both in terms of exploring potential differences in the level of representation abstraction required to predict the two variables and in terms of the part of the item that contains information most relevant to each parameter. This is accomplished by extracting two levels of item representations, *embeddings* and *encodings*, from various parts of the MCQ (answer options only, question only, whole item). Predictions are compared to i) the predictions for each parameter *without* the use of an auxiliary task, and ii) a ZeroR baseline. The results from the transfer learning experiments show the usefulness and limitations of this approach for modeling item parameters with a view to practical scenarios where

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 193–197
July 10, 2020.

we have more data for one parameter. The results for the source of the signal suggest item writing strategies that may be adopted to manipulate specific item parameters.

2 Related Work

The majority of work related to predicting question difficulty has been done in the field of language learning (Huang et al., 2017; Beinborn et al., 2015; Loukina et al., 2016). Some exceptions include estimating difficulty for automatically generated questions by measuring the semantic similarity between the a given question and its associated answer options (Alsubait et al., 2013; Ha and Yaneva, 2018; Kurdi et al., 2016) and measuring the difficulty and discrimination parameters of questions used in e-learning exams (Benedetto et al., 2020). With regards to medical MCQs, previous work has shown modest but statistically significant improvements in predicting difficulty using a combination of linguistic features and embeddings (Ha et al., 2019) as well as predicting the probability that an item meets the difficulty and discriminatory power criteria for use in live exams (Yaneva et al., 2020).

The literature on response time prediction is rather limited and comes mainly from the field of educational testing. The range of predictors that have been explored includes item presentation position (Parshall et al., 1994), item content category (Parshall et al., 1994; Smith, 2000), the presence of a figure (Smith, 2000; Swanson et al., 2001), and item difficulty and discrimination (Halkitis et al., 1996; Smith, 2000). The only text-related feature used in these studies was word count. A more recent study by Baldwin et al. (2020) modeled the response time of medical MCQs using a broad range of linguistic features and embeddings (similar to Yaneva et al. (2019)) and showed that the predicted response times can be used to improve fairness by reducing the time intensity variance of exam forms.

To the best of our knowledge, the use of transfer learning for predicting MCQ parameters has not yet been investigated. The next sections present an initial exploration of this approach for a sample of medical MCQs.

3 Data

The data consists of ≈ 18,000 MCQs from a high-stakes medical licensing exam. An example of an MCQ is presented in Table 1. Let *stem* denote the part of the question that contains the description of the clinical case and let *options* denote the possible answer choices. All items tested medical knowledge and were written by experienced item-writers following a set of guidelines stipulating adherence to a standard structure. All items were administered as (unscored) pretest items for six standard annual cycles between 2010 and 2015 and test-takers had no way of knowing which items were used for scoring and which were being pretested. All examinees were from accredited[1] medical schools in the USA and Canada and were taking the exam for the first time.

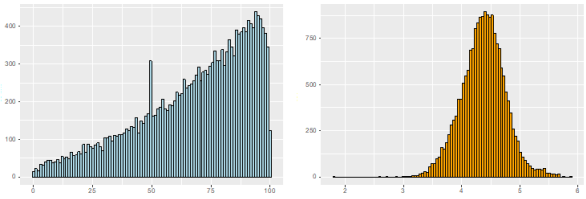

Figure 1: Distribution of the P-value (left) and log Response Time (right) variables

Here, the difficulty of an item is defined by the proportion of its responses that are correct. In the educational testing community this metric is commonly referred to as *P-value*. For example, a P-value of .67 means that the item was answered correctly by 67% of the examinees who saw that item. (Since greater P-values are associated with greater proportions of examinees responding correctly, P-value might be better described as a measure of item easiness than item difficulty.) Response Time is measured in seconds and represents the average amount of time it took all examinees who saw the item to answer it. The distribution of P-values and log Response Times for the data set is presented in Figure 1. The correlation between the two parameters for the set of items is .37.

4 Method

Three types of item text configurations were used as input: i) item stem, ii) item options, and iii) a combination of the stem and options (this combination was used both as a single vector and as two separate vectors). After preprocessing the raw text (tokenization, lemmatization and stopword removal), it was used to train an ELMo (Peters et al., 2018) model[2]. The model was trained with two

[1] Accredited by the Liaison Committee on Medical Education (LCME).

[2] Data pre-processing and feature extraction were implemented using the `PyTorch` and `Allennlp` libraries and the

A 55-year-old woman with small cell carcinoma of the lung is admitted to the hospital to undergo chemotherapy. Six days after treatment is started, she develops a temperature of 38C (100.4F). Physical examination shows no other abnormalities. Laboratory studies show a leukocyte count of 100/mm3 (5% segmented neutrophils and 95% lymphocytes). Which of the following is the most appropriate pharmacotherapy to increase this patient's leukocyte count?
(A) Darbepoetin (B) Dexamethasone
(C) Filgrastim (D) Interferon alfa
(E) Interleukin-2 (IL-2) (F) Leucovorin

Table 1: An example of a practice item

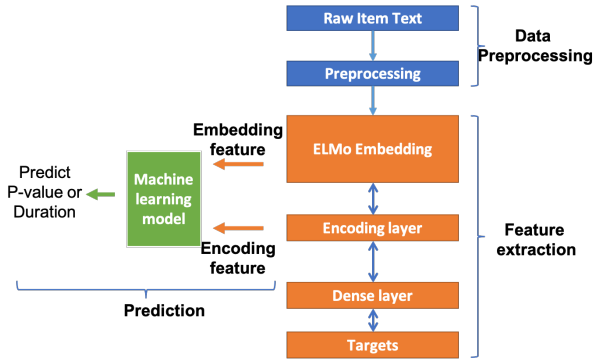

Figure 2: Diagram of the proposed methods.

ELMo types	Number of Parameter	Output dimension	Parameters updating
Small	13.6 million	128	Tuning
Middle	20.8 million	256	Tuning
Original	93.6 million	1024	Freezing

Table 2: ELMo architectures. Parameter tuning was performed for the Small and Middle models. When training the Original ELMo structure, the parameters were frozen (or not updated) because of the memory limitations (6GB) of our NVIDIA Tesla M60 GPU platform.

separate objectives: one was to predict P-value and the other one was to predict Response Time. To learn the sequential information from the ELMo embedding output, an encoding layer was added after the ELMo embedding layers (Figure 2). The encoding layer was constructed using a Bidirectional LSTM network (Graves et al., 2005). This layer allowed the extraction of encoding features, which captured more abstract information than the embeddings alone (the two are later compared). The encoding layer was followed by a dense layer in order to convert the feature vectors to the targets through a non-linear combination of the elements in the feature vectors.

As shown in Table 2, we used three different ELMo configurations (small, middle, and original), each with a different number of parameters. Since the number of parameters of these three ELMo structures was relatively large compared to the size of our item pool, we used the parameters pre-trained on the 1 Billion Word Benchmark (Chelba et al., 2013) as the initialization.

Two modeling approaches were applied. The first approach (Method 1) used the pre-trained ELMo parameters as the initialization and trained on the MCQ data with the aim of predicting the item parameter of interest (either P-value or Response Time). In this scenario, the target variable used in the training procedure was the same as the target variable in the prediction part. The second approach (Method 2) also used the pre-trained ELMo parameters as the initialization but these were updated when training on the auxiliary task. In other words, if the target variable in the prediction part was P-value, then the target variable in the training part was Response Time and vice-versa. Since we are also interested in understanding the effects of different levels of abstraction on parameter prediction (as captured by the embeddings and encodings), we used linear regression (LR) to predict the item characteristics using the extracted features as input. The training set, the validation set and the testing set consisted of 12,000 samples, 3,000 samples, and 3,000 samples, respectively.

5 Results and Discussion

The results for the experiments are presented in Table 3. As can be seen, the models achieved a slight but significant RMSE decrease compared to the ZeroR baseline. In addition, Method 2 significantly improved the prediction of the Response Time variable (when predicting P-value is used as an auxiliary task) but this was not the case the other way around (predicting P-value with Response Time as an auxiliary task). A possible explanation for this result is the fact that the models were much better at predicting the Response Time component

prediction part was implemented using the `scikit-learn` library. The NVIDIA Tesla M60 GPU was used to accelerate the model training.

		P-value (M1)		Resp. Time (M1)		P-value (M2)		Resp. Time (M2)	
ELMo	**Item component**	**Embed**	**Encod**	**Embed**	**Encod**	**Embed**	**Encod**	**Embed**	**Encod**
Original	Stem	23.60	**23.32**	0.31	0.33	23.62	23.33	**0.31**	0.33
Original	Answer options	23.61	23.35	0.35	0.35	23.63	23.29	0.35	0.35
Original	Full Item	23.55	23.43	0.31	0.33	23.60	23.34	0.31	0.34
Original	Stem + Options	23.71	23.40	0.32	0.32	24.24	23.27	0.32	**0.33**
Average		23.61	23.38	0.32	0.33	23.77	23.31	0.32	0.34
Small	Stem	23.74	23.67	0.31	0.30	23.16*	23.20*	0.33	0.33
Small	Answer options	23.48	23.68	0.35	0.34	23.23*	23.31*	0.36	0.36
Small	Full Item	27.64	NA	0.31	0.30	23.20	23.23	0.38	0.70
Small	Stem + Options	23.71	23.71	0.30	**0.29**	**23.04***	23.21*	0.34	0.33
Average		24.64	23.69	0.32	0.31	23.16	23.24	0.35	0.43
Middle	Stem	23.54	23.73	0.31	0.30	23.32	**23.16***	0.32	0.36
Middle	Answer options	24.90	23.74	0.35	0.35	23.67	23.38*	0.37	0.37
Middle	Full Item	**23.45**	23.65	**0.30**	0.30	23.39	23.22*	0.32	0.33
Middle	Stem + Options	24.76	23.95	0.31	0.30	23.82	23.24*	0.33	0.37
Average		24.16	23.77	0.32	0.31	23.55	23.25	0.34	0.36
Total aver.		24.17	23.60	0.32	0.32	23.49	23.26	0.34	0.38
Baseline									
ZeroR		23.97		0.35					

Table 3: Results for P-value and Response Time using Method 1 (columns 3-4) and Method 2 (columns 5-6). The values represent the Root Mean Squared Error (RMSE) for each model obtained using linear regression. Values marked with * represent cases, where the use of Method 2 has resulted in a statistically significant improvement compared to Method 1 (95% Confidence Intervals). The best result in each column is marked in red.

compared to the ZeroR baseline and this knowledge successfully transferred into improving the P-value prediction. The gains in predicting the P-value on the other hand were much more modest, which may explain why they did not contribute to the prediction of Response Time. Another possible explanation could be that P-values were highly skewed whereas Response Times were normally distributed. It could be that the normalized distribution of the Response Time variable facilitates learning of better representations compared to the skewed distribution of the P-value variable. A direction for future work is to test this by normalizing both distributions.

Not all parts of the item were equally important for predicting the two parameters. Signal from the stem alone provided the best results for the P-value variable in Method 1 (23.32) and when P-value was used as an auxiliary task for predicting Response Time (0.31) in Method 2 (i.e., adding information from the answer options did not improve the result). By contrast, signal from the full item outperformed other configurations when the Response Time was predicted using Method 1 (0.29) and when Response Time was used as an auxiliary task for predicting the P-value (23.04). Therefore, the stem contained signal that was most relevant to the P-value variable, while the Response Time was best predicted using information from the entire item. This suggests that deliberating between the different answer options and reading the stem all have effects on the Response Time. However, the difficulty of the clinical case presented in the stem seems to have a stronger relation to the P-value than the difficulty attributed to choosing between the answer options. Using the stem and options content as two predictors (Stem + Options) had no significant effects but, on average, provided slightly more accurate results than the single predictor (Full Item). Finally, no clear pattern emerged with regards to the predictive utility of using embeddings vs. encodings or the embedding dimensions and weight tuning produced by training the three ELMo models (Small, Middle and Original).

These results represent a first step towards the exploration of transfer learning for item parameter prediction and may have implications for both parameter modeling and item writing.

6 Conclusion

This study investigated the use of transfer learning for predicting difficulty and Response Times for clinical MCQs. Both parameters were predicted with a small but statistically significant improvement over ZeroR. This prediction was further improved for P-value by using transfer learning. It was also shown that the item stem contained signal that was most relevant to the P-value variable, while the Response Time was best predicted using information from the entire item.

References

Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2013. A similarity-based theory of controlling mcq difficulty. In *e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on*, pages 283–288. IEEE.

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2020. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. R2de: a nlp approach to estimating irt parameters of newly generated questions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 412–421.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*, pages 799–804. Springer.

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 389–398.

Le An Ha, Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20.

Perry N Halkitis et al. 1996. Estimating testing time: The effects of item characteristics on response latency.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *AAAI*, pages 1352–1359.

Ghader Kurdi, Bijan Parsia, and Uli Sattler. 2016. An experimental evaluation of automatically generated multiple choice questions from ontologies. In *OWL: Experiences And directions–reasoner evaluation*, pages 24–39. Springer.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253.

Cynthia G Parshall et al. 1994. Response latency: An investigation into determinants of item-level timing.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Russell Winsor Smith. 2000. An exploratory analysis of item parameters and characteristics that influence item level response time.

David B Swanson, Susan M Case, Douglas R Ripkey, Brian E Clauser, and Matthew C Holtman. 2001. Relationships among item characteristics, examine characteristics, and response times on usmle step 1. *Academic Medicine*, 76(10):S114–S116.

Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20.

Victoria Yaneva, Le An Ha, Peter Baldwin, and Janet Mee. 2020. Predicting item survival for multiple choice questions in a high-stakesmedical exam. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , Marseille, 11–16 May 2020*, page 6814-6820.