Accurate Modelling of Language Learning Tasks and Students Using Representations of Grammatical Proficiency

Ahmed H. Zaidi
Computer Laboratory
University of Cambridge
ahz22@cl.cam.ac.uk

Andrew Caines
Computer Laboratory
University of Cambridge
apc38@cam.ac.uk

Christopher Davis
Computer Laboratory
University of Cambridge
ccd38@cam.ac.uk

Russell Moore
Computer Laboratory
University of Cambridge
rjm49@cam.ac.uk

Paula Buttery
Computer Laboratory
University of Cambridge
pjb48@cam.ac.uk

Andrew Rice
Computer Laboratory
University of Cambridge
acr31@cam.ac.uk

ABSTRACT

Adaptive learning systems aim to learn the relationship between curriculum content and students in order to optimise a student's learning process. One form of such a system is content recommendation in which the system attempts to predict the most suitable content to next present to the student. In order to develop such a system, we must learn reliable representations of the curriculum content and the student. We consider this in the context of foreign language learning and present a novel neural network architecture to learn such representations. We also show that by incorporating grammatical error distributions as a feature in our neural architecture, we can substantially improve the quality of our representations. Different types of grammatical errors are automatically detected in essays submitted by students to an online learning platform. We evaluate our model and representations by predicting student scores and grammatical error distributions on unseen language tasks.

1. INTRODUCTION

In general the adaptive learning approach has been shown to lead to improved learning outcomes for student users of educational platforms [4, 8, 10]. However, there remains a question of what is the best methodology to construct representations for students and tasks. Previous approaches manually engineer features to construct representations [7]. These features are usually tuples of a knowledge component (e.g. differentiation, fractions in the case of maths) and student outcome (i.e. whether or not the student demonstrated understanding for that knowledge component through completing the task). A task may contain multiple knowledge components. Whilst this approach is highly interpretable, in the domain of language learning, it is difficult to clearly divide the tasks into knowledge components. Furthermore, in the recently popular paradigm of deep learning, we have seen that training representations through neural networks have yielded state-of-the-art results in the space of image recognition, and various natural language tasks.

Motivated by this, we propose a methodology of automatically developing high quality representations of students and tasks in a language learning context. Having reliable student and task representations in place facilitates work on downstream tasks such as curriculum learning and recommender systems for language learning.

Representations are derived from a novel neural architecture and real student data collected through the Write & Improve[1] (W&I) assessment and feedback platform for learners of the English language. [13].

Our best-performing model incorporates grammatical error distributions detected by ERRANT [3] as a feature and achieves mean squared error (MSE) of 1.195 on score prediction, an absolute value of 1.093 on a scoring scale of 0-13.

2. WRITE & IMPROVE

On W&I, students are prompted to input a short text of at least 25 words in response to a given question. Once they have completed the task, the W&I automarker assigns each text an integer score s between 0 and 13. The system also automatically provides a grade on the CEFR scale[2] along with feedback on grammatical errors detected in the text. Table 1 outlines how integer scores are mapped to the CEFR scale. Currently all users of W&I move through the curricu-

Table 1: Student scores mapped to CEFR levels

A1	A2	B1	B2	C1	C2
1-2	3-4	5-6	7-8	9-10	11-13

[1] https://writeandimprove.com

[2] The Common European Framework of Reference for Languages

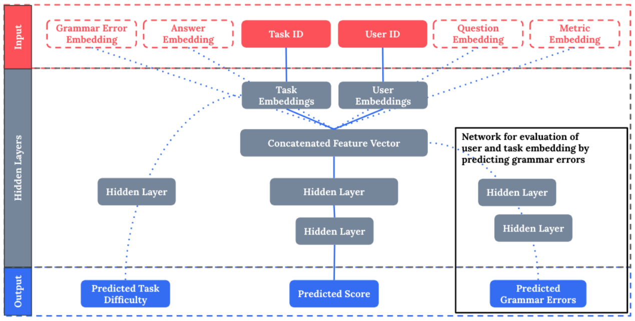

Figure 1: Task score prediction system architecture. Dotted lines and boxes are optional features and network connections.

lum in an unguided and independent fashion. An intelligent tutoring system would instead guide students from task to task in order to personalise their learning experience and improve their level of performance. In order to do that we must learn reliable student and task representations.

We obtained application logs of user activity from the past two years – a total of 3+ million essay submissions by 300k account holders. We filtered the data for users who had submitted at least 10 submissions. We also had a record of the questions ('prompts') users responded to and the scores assigned to their texts by W&I's auto-marker.

3. LEARNING STUDENT AND TASK REPRESENTATIONS

Our primary goal was to predict student scores on a given language learning task based on our representations of students and tasks in W&I. Secondary to that, we check the quality of our student representations by predicting the grammar error distribution of a given student-task tuple. In what follows we describe the data, evaluation metrics and models used in this work.

3.1 Model Architecture

The architecture of our neural system can be seen in Figure 1. The neural network takes as an input a user id u and task id t which are taken as indices in the user embedding layer U and task embedding layer T respectively. $u \in N_u$ where N_u is the number of unique users in the W&I dataset. $t \in N_t$ where N_t is the number of unique tasks in the W&I dataset.

We optimise our system and learn a user embedding matrix U and task embedding matrix T by minimising the mean squared error (MSE) of our predicted score s and the target score $\hat{s}$.

We introduce an auxiliary objective to predict the difficulty β of each task t, referenced as t_β. The ground-truth labels for task difficulty (beginner, intermediate, advanced) are obtained from the meta-data of each task in the W&I dataset.

3.2 Feature Set

In addition to the score s, the W&I dataset contains prompts and answers in natural language as well as metrics on whether submission k is the highest scoring submission by user u. We incorporate these additional features into the architecture of the model in order to evaluate their impact on the quality of user and task embeddings.

3.2.1 Answer and Question Embedding

We obtain a vectorised form of each student response and question using 300-dimension word2vec embeddings[3] pre-trained on the Google News corpus [5]. Our embeddings are an additive compositional model where the final embedding is a sum of every word in the question or answer. Whilst this model is not state-of-the-art for distributional semantics, Mitchell & Lapata [6] show that the additive model can yield results comparable to more sophisticated models.

3.2.2 Metric embedding

The metric embedding is a 2-dimensional vector. The first dimension is a binary value for whether the score for the submission was the highest score on task t for user u. The second dimension is a binary value for whether the score for the submission was the highest score across all W&I tasks for user u.

3.2.3 Grammar error embedding

A student's grammatical proficiency plays a vital role in determining how well they perform on a particular task. As we do not know of any system that identifies appropriate use of grammar, we focused on understanding what grammatical structures the student struggles with. This was done by running ERRANT [3], an automated error detection and correction system, in order to identify grammatical errors in the student's essay.

For each submission k, we constructed a 47-dimensional vector, one dimension for each of the error types observed in the W&I dataset. Each dimension stored the number of times that error type appeared in the student's essay submission.

$$< e_k >=< f_k^1, f_k^2, \ldots, f_k^{47} > \tag{1}$$

– where e_k is the grammar error embedding e for submission k, and f_k^n is the frequency of errors for error type n in submission k.

3.3 Mean score baseline

Our baseline system for predicting s for user u on task t is to calculate the mean of observed scores by all users for that task. We refer to this baseline as MEAN_SCORE.

3.4 Evaluation

We identify two approaches to evaluating our system and the quality of our learned user and task representations: 1) score prediction; and 2) grammar error prediction.

[3] A word2vec embedding is a $1 \times x$ dimensional dense vector that represents a word semantically.

Table 2: Score prediction (MSE) and grammar embedding prediction (cosine) results for the top 8 best performing feature combinations (error: grammar error embedding; ques: question embedding; ans: answer embedding; metric: metric embedding).

Model	MSE	Cosine
MEAN_SCORE (baseline)	1.913	-
error+ques+ans+metric	2.254	-0.385
ques+metric	1.942	-0.402
ans+metric	1.951	-0.414
error+metric	**1.350**	**-0.426**
ques	2.028	-0.403
ans	2.014	-0.412
error	1.761	-0.410
metric	1.907	-0.393

3.4.1 Evaluation of score predictions

To evaluate the performance of score prediction we use mean squared error (MSE) in common with other works in this field, using global computation where all data points are treated equally [9].

3.4.2 Evaluation of grammar embedding predictions

In order to further evaluate the quality of the learned user and task representations, we also introduce an additional evaluation task of predicting the distribution of grammar errors for a user u on a task t.

This was done by building a network that takes as an input the user $\vec{u}$ and task $\vec{t}$ from the pre-trained embedding U and T and predicts the grammar embedding $\vec{g}$. Our dataset for grammar error prediction was created by extracting the last submission k of every user u. This was to ensure that the system is predicting the distribution of errors for the users at their most recent knowledge state.

We optimise our system by minimising the cosine proximity of the predicted grammar vector $\vec{g}$ and the target grammar vector $\hat{\vec{g}}$.

4. RESULTS

Table 2 summarises the results of our system. We compare the effectiveness of various features in the prediction of a user's score s on a task t which is evaluated by MSE. We include the top 8 MSE values on the score prediction system and their corresponding cosine value from the grammar error prediction model.

We find that incorporating question and answer embeddings do not provide any performance improvement in terms of MSE beyond the baseline model. The metric embedding provides marginally better results than the baseline with an MSE of 1.907. The grammatical error embedding provides substantial improvements beyond both the baseline and the metric embedding with an error of 1.761. The best performing system incorporates both grammatical error embedding and metric embedding, reducing the MSE to 1.350.

Table 3: Performance across various student and task representations sizes (N_h)

Model	N_h	MSE	Cosine
error+metric	3	1.350	-0.426
error+metric	5	1.297	-0.431
error+metric	16	1.245	-0.415
error+metric	32	**1.195**	**-0.433**

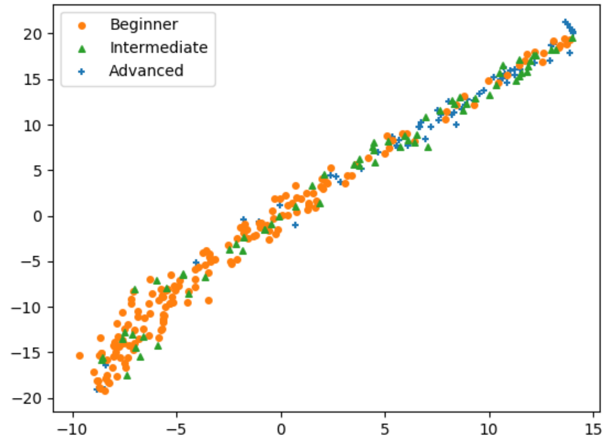

Figure 2: t-SNE of 300 randomly sampled student representations classified by different levels of proficiency

Table 3 shows the model that provides the lowest cosine proximity to the target grammatical error vector (i.e. best system) was error+metric, which is consistent with the lowest MSE for the score prediction system.

In order to interpret the relevance of cosine proximity we conducted a Pearson's correlation test between the MSE values from the score prediction system and the cosine proximity scores from the grammar error prediction system. The results show a 0.7883 Pearson's correlation with a p-value of 0.0201 which is statistically significant at $\alpha < 0.05$.

Figure 2 shows a t-SNE [12] of 300 randomly sampled student representations learned by our best performing score prediction system. The students are classified by their proficiency which has been determined by observing the most frequent task level attempted in their five most recent submissions. Qualitatively, the results from the plot are promising as the advanced and intermediate users, whilst present throughout the plot, are more concentrated towards the top right (higher level of language proficiency). Beginner students, on the other hand, are more concentrated in the bottom left. This suggests that the embeddings constructed from our model provide context on the language abilities of the student.

5. DISCUSSION

The results in Table 2 show that incorporating grammar error embeddings provides a reliable signal to learn wellformed student and task representations. Furthermore, Table 3 identifies the optimal size for student and task representations by training the system using various configurations and evaluating both the MSE and cosine. Larger embedding size performed better than the smaller embedding sizes up to our experimental maximum of 32 dimensions. However, making the embedding size too large would result in what is known as 'overcomplete' which in turn causes the model to simply memorise the correct response instead of learning discriminative features [2].

In real terms, an MSE of 1.195 represents a root mean squared error of 1.093 on a scale of 0 to 13. This means that on average we stay within the bounds of a CEFR level when predicting student proficiency which seems sufficiently robust for real world application.

Grammar errors highlight the weaknesses of the student as opposed to their strengths. Therefore, instead of learning the upper-bound of a student's ability, we are learning the features for the lower-bound. The results of the model also suggest that there is a correlation between the types of errors students make on task t and the score they achieve on said task. This enables the model to learn latent features within the student and task representations which in turn can be used to reliably predict the student's score on a future unseen task.

The importance and value of the signal provided by grammar errors in determining student ability and thus creating quality representations can be further highlighted by Figure 3. The bar-chart shows a comparison between beginner and intermediate students, where the values in x-axis are the various error types in ERRANT and the values for the y-axis are the normalised difference of the frequency for each error type between the two groups of students (positive bars indicate greater frequency of that error type for intermediate students). We can observe that certain errors such as `M:VERB:TENSE` (highlighted in orange) are more frequent with intermediate students. This is not surprising as beginner students tend not to experiment with verb tenses but rather focus on using verb tenses that they are comfortable with. Intermediate students are more likely to learn verb conjugation rules and over-regularise to introduce variation in sentence structure. However, over-regularisation usually results in increased number of verb tense errors [11, 1]. This is then corrected once students reach an advanced level of proficiency where they can account for the irregular verb tenses.

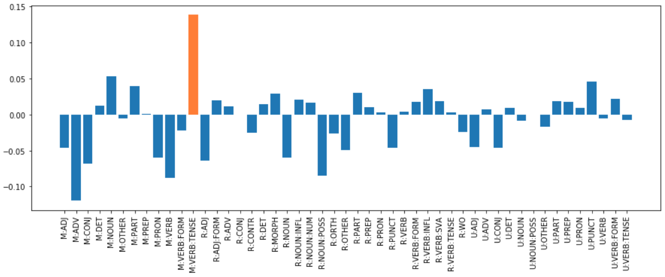

Figure 3: A bar-chart showing a comparison of errors between beginner and intermediate students.

6. REFERENCES

[1] K. Bardovi-Harlig. Tense and aspect in second language acquisition: form, meaning, and use. *Language Learning: A Journal of Research in Language Studies*, 50:1, 2000.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] C. Bryant, M. Felice, and T. Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017.

[4] R. V. Lindsey, J. D. Shroyer, H. Pashler, and M. C. Mozer. Improving studentsâĂŹ long-term knowledge retention through personalized review. *Psychological Science*, 25(3):639–647, 2014.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[6] J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

[7] S. Montero, A. Arora, S. Kelly, B. Milne, and M. Mozer. Does deep knowledge tracing model interactions among skills? In *Proceedings of the 11th International Conference on Educational Data Mining (EDM)*, 2018.

[8] A. S. Najar, A. Mitrovic, and B. M. McLaren. Learning with intelligent tutors and worked examples: selecting learning activities adaptively leads to better learning outcomes than a fixed curriculum. *User Modeling and User-Adapted Interaction*, 26:459–491, 2016.

[9] R. Pelánek. The details matter: methodological nuances in the evaluation of student models. *User Modeling and User-Adapted Interaction*, 28:207–235, 2018.

[10] Y. Rosen, I. Rushkin, R. Rubin, L. Munson, A. Ang, G. Weber, G. Lopez, and D. Tingley. The effects of adaptive learning in a massive open online course on learners' skill development. In *Proceedings of Learning @ Scale*, 2018.

[11] D. E. Rumelhart and J. L. McClelland. On learning the past tenses of English verbs. In D. E. Rumelhart, J. L. McClelland, and PDP Research Group, editors, *Parallel distributed processing: explorations in the microstructure of cognition, vol. 2*. Cambridge, MA: MIT Press, 1986.

[12] L. J. P. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:579–2605, 2008.

[13] H. Yannakoudakis, Ø. E. Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31:251–267, 2018.