Question Generation for Adaptive Education

Megha Srivastava
Stanford University
megha@cs.stanford.edu

Noah Goodman
Stanford University
ngoodman@stanford.edu

Abstract

Intelligent and adaptive online education systems aim to make high-quality education available for a diverse range of students. However, existing systems usually depend on a pool of hand-made questions, limiting how fine-grained and open-ended they can be in adapting to individual students. We explore targeted question generation as a controllable sequence generation task. We first show how to fine-tune pre-trained language models for deep knowledge tracing (LM-KT). This model accurately predicts the probability of a student answering a question correctly, and generalizes to questions not seen in training. We then use LM-KT to specify the objective and data for training a model to generate questions conditioned on the student and target difficulty. Our results show we succeed at generating novel, well-calibrated language translation questions for second language learners from a real online education platform.

1 Introduction

Online education platforms can increase the accessibility of educational resources around the world. However, achieving equitable outcomes across diverse learning needs benefits from systems that are adaptive and individualized to each student (Doroudi and Brunskill, 2019). Traditionally, adaptive education methods involve planning over a pool of pre-made questions (Atkinson, 1972; Hunziker et al., 2018). These are naturally limited by the diversity and coverage of the pool, as well as the scaling capacity of curriculum planning algorithms. Recent approaches, such as procedural generation for personalized programming games (Valls-Vargas et al., 2017), are limited to well-specified small domains. We address these limitations by leveraging recent success in deep generative models, in particular language models (LMs).

Many educational activities involve sequential data, such as language translation, reading compre-
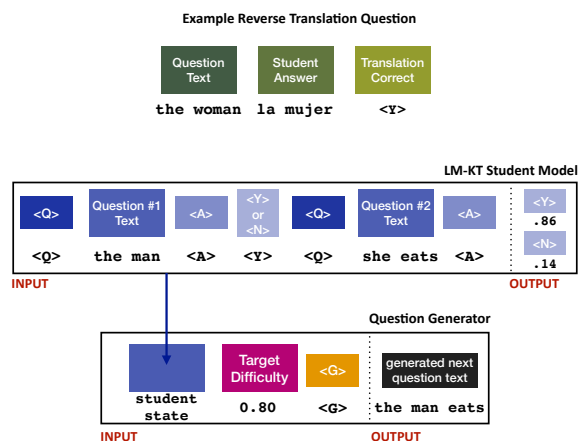

Figure 1: Example input and outputs for our LM-based knowledge tracing model (middle) and question generation model (bottom) for an online reverse language translation task (top). A question in this task consists of a target phrase for the student, in this case a Spanish learner, to translate (e.g. "the woman").

hension, algebra, and deductive logic. Meanwhile, pre-trained LMs can effectively handle sequences from a wide range of modalities (Madani et al., 2020; Polu and Sutskever, 2020). In this work, we focus on natural language sequences, where recent progress in language modeling has shown great success at capturing abstract properties of language (Hewitt and Manning, 2019; Liu et al., 2019). Specifically, we show how pre-trained LMs can be easily leveraged to adaptively generate questions for a given student and target difficulty in a *reverse translation task*, using difficulty at answering questions as a proxy for more complex future learning objectives.

We introduce an LM-based knowledge tracing model (LM-KT) to predict students' difficulty on novel questions (e.g. target phrases to translate). We show that LM-KT is well-calibrated, allowing us to pose the learning problem for the question generator: given a student state, generate a question that will achieve a target difficulty, according

692

to LM-KT. We evaluate both LM-KT and question generation models on real users and responses from Duolingo[1], a popular online second-language learning platform.

2   Background & Related Works

There exists a rich body of work on precisely modeling student "ability" and learning. For example, Item Response Theory (IRT) seeks to model individual student ability based on their responses to different questions, creating a strong factorization between students and test items (Lord, 1980; Hambelton and Jodoin, 2003). Meanwhile, Computer Adaptive Testing (CAT) techniques are used to determine a fixed student ability as quickly as possible by selecting test items based on information utility (Weiss and Kingsbury, 1984; Thissen and Mislevy, 2000; Settles et al., 2020). However, these methods, which have been used to develop efficient standardized tests, do not necessarily optimize a student's *learning* experience (Mu et al., 2018). We instead focus on tracking each student's evolving knowledge, choosing questions to target difficulty.

Knowledge Tracing (KT) seeks to model a student's knowledge state from their answer history in order to help individualize exercise sequences (Corbett and Anderson, 1995). This draws inspiration from traditional education curriculum practices, such as distributed spacing of vocabulary (Bloom and Shuell, 1981) and mixed review in mathematics (Rohrer, 2009). To address simplifying assumptions in earlier KT approaches, such as discrete knowledge representations, Piech et al. (2015) introduced Deep Knowledge Tracing (DKT), which uses RNNs to enable more complex knowledge representations for students. Recently, SAINT+ (Shin et al., 2020) showed state-of-the-art performance on the popular EdNet KT task using a Transformer model to capture temporal information across activities, motivating our use of Transformer LMs.

Controllable Text Generation   aims to steer LMs towards desired attributes. Examples include using reinforcement learning to control quality metrics (Ranzato et al., 2016), adjusting sampling weights to control for poetry style (Ghazvininejad et al., 2017), and learning to condition on valence or domain-specific codes (Keskar et al., 2019; Peng et al., 2018). To the best of our knowledge, we are

the first to use controllable generation in an education context with real student interaction data.

3   Method

Given any autoregressive language model (e.g. GPT-2 (Radford et al., 2019), we can fine-tune a **LM-KT model** ($p_{\theta_{KT}}$) to predict whether an individual student will correctly answer the next question. If this model has well-calibrated uncertainty, we can use its predicted probability of a correct answer as a proxy for the difficulty of a question to a student. We then train a **question generation model** ($p_{\theta_{QG}}$) to generate a new question conditioned on a student and desired target difficulty.

Question Representation   Unlike standard DKT, which treats questions as IDs or simple hand-crafted features, we represent questions fully in text (e.g. "she eats" in Figure 1). This is a key contribution of our work, required by our eventual goal of *generating* questions in text, and allows the model to leverage similarity across linguistic features. We thus represent a question q as a sequence of words, with prefix and suffix tokens:

$$q_i = \quad \texttt{<Q>} \quad w_1^i \quad w_2^i \quad w_3^i \quad ... \quad w_n^i \quad \texttt{<A>}$$

Student State   We represent a student as a temporally-evolving sequence of questions and their responses. As in much previous KT work, we represent the student response as simply correct/incorrect, with special tokens $\texttt{<Y>}$ and $\texttt{<N>}$. A student's current state is thus represented as a sequence of all past question and response pairs:

$$s_j = \quad q_1^j \; a_1^j \; q_2^j \; a_2^j \; ... \; q_m^j \; a_m^j \;, a_i \in \{\texttt{<Y>},\texttt{<N>}\}$$

LM-KT   Given the sequential nature of student learning over time, we can easily frame knowledge tracing as an autoregressive language modeling task. Given a dataset D of students $s_1, s_2, ..., s_{|D|}$, we employ the standard training objective of finding the parameters θ_{KT} that minimizes

$$\mathscr{L}_{KT} = -\sum_{i=1}^{|D|} \sum_{t=1}^{|\mathbf{x}^{(i)}|} \log p_{\theta_{KT}}(x_t^{(i)}|x_{<t}^{(i)}) \quad (1)$$

where $\mathbf{x}^{(j)} = (x_1^{(j)}, ..., x_{|\mathbf{x}|}^{(j)})$ is the entire sequence tokens corresponding to student s_j, consisting of all their past questions and answers. Using the softmax output of the LM-KT model ($p_{\theta_{KT}}$), we estimate a student's (inverse) difficulty in answering a specific question as $d_{qs} = p_{\theta_{KT}}(\texttt{<Y>}|s, q)$. We find that $p_{\theta_{KT}}$ is well-calibrated (Section 4.2), yielding a good proxy for the true question difficulty.

Question Generation We frame question generation as finetuning a *new* autoregressive LM. Given random samples of students and questions from a held-out set not used to train LM-KT, we can construct a new dataset D' consisting of s_i d_i<G> q_i sequences, where <G> is a special generation token and $d_i = p_{\theta_{KT}}(\text{<Y>}|s_i, q_i)$ is the continuous difficulty value assigned by LM-KT. We learn a linear layer to map the continuous input difficulty into a *difficulty control vector* c_d of dimension matching the LM word-embeddings, which we append to the token embeddings. Unlike LM-KT, we train our question generation model $p_{\theta_{QG}}$ to minimize the loss only on the question text, which only appears *after* the <G> token. If t_g is the token index of <G>, then our modified loss is:

$$\mathscr{L}_{QG} = -\sum_{i=1}^{|D'|} \sum_{t=t_g+1}^{|\mathbf{x}^{(i)}|} \log p_{\theta_{QG}}(x_t^{(i)}|x_{<t}^{(i)}) \quad (2)$$

where sequence $\mathbf{x}^{(j)}$ contains the full s_j d_j<G>q_j sequence. At test time, we generate tokens $w_1...w_n$ conditioned on the s_j d_j <G> prefix.

4 Experiments

Our method generalizes to any education activity that can be represented with text sequences. Due to the availability of real student learning data, we focus on a *reverse language translation* task, where a student translates phrases from their native language (e.g. English, "she eats") to the second language they are learning (e.g. Spanish, "ella come").

4.1 Experimental Details

We use the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (Settles et al., 2018) dataset, which contains questions and responses for Duolingo users over the first 30 days of learning a second language. While the original task's goal was to identify *token*-level mistakes, we collapse these errors into binary (correct / incorrect) per-question labels. We use the provided train/dev/test splits for users learning Spanish and French. We create separate held-out sets from the test set to evaluate the LM-KT and question generation models. For both models, we finetune separate GPT-2 (Radford et al., 2019) models. While we sample from a held-out set of student states and questions to train the question generation model, in principle questions can come from any source text

Model *(Spanish)*	AUC (seen)	AUC (unseen)
LM-KT	**0.75** ±.0001	**0.76** ±.001
Standard DKT	0.72 ±.0001	0.70 ±.001
Question Only	0.67 ±.0001	0.58 ±.002
Model *(French)*	AUC (seen)	AUC (unseen)
LM-KT	**0.73** ±.0002	**0.71** ±.002
Standard DKT	0.70 ±.0001	0.65 ±.002
Question Only	0.65 ±.0002	0.62 ±.001

Table 1: LM-KT improves AUC for both questions in the Duolingo test set that were seen during training (for other students) *and* novel questions, over Standard DKT with Question IDs and question-only baselines. Errors are 95% CIs.
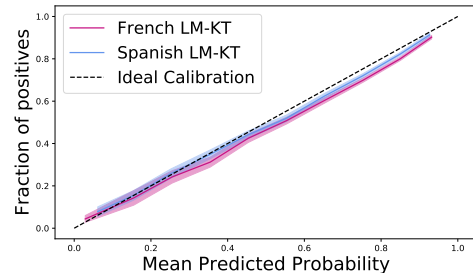

Figure 2: Both LM-KT models are well calibrated, but the French model is slightly more overconfident. Filled area shows bootstrap (n=1000) standard deviation.

domain. Further experiment details are in the Appendix, and source code can be found at: `https://github.com/meghabyte/acl2021-education`.

4.2 Results: Student Modeling

We evaluate LM-KT two ways: first, its ability to predict if an individual student will answer a novel question correctly on a held-out test set of real Duolingo student responses. Second, how well-calibrated these predictions are, which is crucial to our later use of LM-KT for question generation.

Table 1 compares AUC-ROC on a held-out test set for our LM-KT model with standard DKT, which uses question IDs instead of text, and a baseline that ignores the student state, only using the question text representation. This question only baseline would perform well if the Duolingo dataset largely consisted of universally "easy" and "difficult" questions, independent of individual student. Our results show that incorporating the student state is crucial for accurately predicting Duolingo user responses, and including question text also leads to a significant improvement. LM-KT outperforms Standard DKT especially on novel questions—a necessary generalization ability for generation.

Finally, we measure the calibration of our LM-KT models for both Spanish and French (from En-

glish) learners, which is the crucial property for our downstream generation task. We bin our test data by predicted question difficulty, and plot the fraction of true correct answers in each bin. Figure 2 shows that LM-KT is well-calibrated, for both Spanish and French, meaning the predicted difficulty matches the empirically observed proportion of correct answers.

4.3 Results: Question Generation

We evaluate four different aspects of our question generation model: (i) successful control for difficulty, (ii) novelty, (iii) fluency, and (iv) latency.

Difficulty Control To explore whether our question generation model indeed depends on target difficulty and the individual student, we first measure the model's perplexity on a held-out test set of Duolingo questions, compared to permutation baselines. Table 2 (top) shows that perplexity is lower for true student / target difficulty inputs than when either or both of these are permuted. The target difficulty values in this analysis were defined by the LM-DKT model. We can remove this dependence by using the actual student responses from Duolingo: we set the target difficulty to 1 if the student was correct and 0 otherwise. Table 2 (bottom) shows our model prefers questions paired with these "true correctness" targets than paired with random ones.

To evaluate how well our generation model achieves target difficulties, we take 15 unseen students and generate 30 questions for each of 9 input difficulties (0.1-0.9). We then use LM-KT (a well-calibrated proxy for true difficulty) to measure the difficulty of these generated questions for each student. Figure 3 shows that we are able to achieve fine-grained control over target difficulty for both Spanish and French students, with an average Root-Mean Squared Error (RMSE) of **.052** across all students and target difficulties. Adding a sampling penalty (Keskar et al., 2019) increases the variance in difficulty (RMSE .062) in exchange for more novel and diverse questions, as discussed next.

Novelty and Fluency By leveraging a pre-trained language model's ability to manipulate structure, we can generate novel questions not present in the entire Duolingo question set (See Table 3). Across 4,050 questions generated for Spanish learners, we found that with a repetition penalty (Keskar et al., 2019), around 43% of all questions, and 66% of high difficulty ($d = 0.1$)
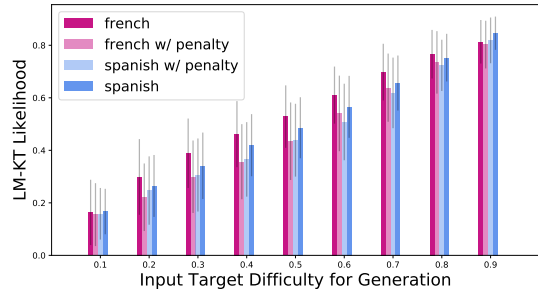

Figure 3: For a random selection of 15 students, our question generator successfully controls for difficulty across a range of 9 target values, evaluated by the LM-KT model. Error bars show standard deviation.

questions, were novel [2]. For French learners, 48% of all and 55% of high difficulty ($d = 0.1$) questions were novel. However, around 3% of generated sentences were judged to be non-fluent [3], although most were still able to be translated (e.g."if i eat some baguettes it breaks."). Without a sampling penalty, the proportion of novel questions drops to about 11 % of questions for French learners (6 % for Spanish learners), yet with far fewer non-fluent examples. Further details and examples of novel and non-fluent generated questions for both Spanish and French learners, are in the Appendix.

Ablation Type	ppl Spanish	ppl French
LM-DKT Likelihood (0 - 1)		
Ground Truth	**4.33** ±0.20	**3.86** ±0.09
Permute Student	6.73 ±0.24	5.11 ±0.41
Permute Difficulty	12.5 ±1.01	7.66 ±0.33
Permute Both	13.1 ±0.43	7.87 ±0.26
Real Student Answers (0 or 1)		
Ground Truth	**17.7** ±1.3	**9.49** ±.20
Permute Student	19.75 ±0.43	10.56 ±.60
Permute Difficulty	30.6 ±2.17	13.5 ±0.60
Permute Both	31.3 ±1.49	13.8 ±0.43

Table 2: Perplexity of the question generation model over a held-out evaluation set with ablations.

Latency Positive student experience in online education requires low latency. In about four seconds, our model can generate 30 questions close to a target difficulty. An alternative to question generation is to rank questions from a preexisting pool, according to a target difficulty objective. We compare the quality (RMSE in achieving target difficulty) of the top 30 questions in a pool against the run-time

---

[2] The CTRL penalty discounts the scores of previously generated tokens, with the HuggingFace Transformers library (Wolf et al., 2020) implementation including tokens provided as part of the prompt. In our setting, this effectively penalizes for generating questions already seen by the student.

[3] We use the language-check Python tool to verify grammar https://pypi.org/project/language-check/.

Difficulty: 0.1 (very hard)	0.9 (very easy)
you write letters.	*spoon or tea*
i know about that book.	socks
she reads your letters.	good morning!
those ducks drink water.	a horse
he mixes coffee with water.	*oil against salt*
Difficulty: 0.3 (hard)	**0.7 (easy)**
you drink juice or water	*what dream?*
you drink water	saturday and sunday
accordingly he does it	until tomorrow
can we as a band?	*yes, it is possible!*
during the night	me too

Table 3: Example questions generated by our model for a Spanish learner. *Italic questions* are novel, and do not exist in the Duolingo dataset.

required to rank all questions in the pool, varying its size (Figure 4). On one NVIDIA Titan XP GPU, we find that, averaged across all target difficulties, our question generation model takes half the time to achieve the same quality as pool selection. The gap increases when trying to sample harder questions (d <0.5) – even a pool size of 1000 does not have sufficient difficult questions, likely due to a skew in the Duolingo question set. Additional controls, such as for style or topic, can easily be combined with our generation method, but would make pool selection exponentially more complex.
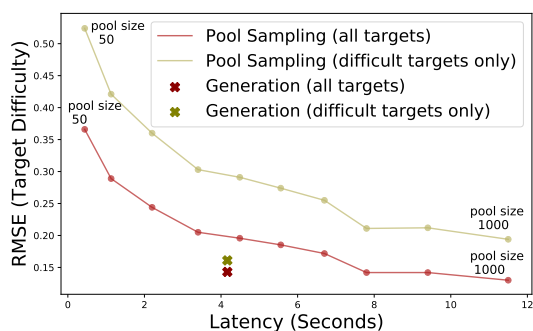

Figure 4: Pool selection (for one student) suffers worse question quality vs. latency trade-off than question generation, especially for sampling difficult questions.

5 Conclusion

Our work is a first step toward showing that sequence-based models combined with domain knowledge, such as pre-trained LMs, can be leveraged for adaptive learning tasks. We show how to use modern LMs to generate novel reverse-translation questions that achieve a target difficulty, allowing adaptive education methods to expand beyond limited question pools.

Limitations of our approach include the compute constraints of large LMs and training data availability. More detailed student data will be crucial to

future model development. For instance, while most publicly available education datasets do not include the full student responses (e.g. full translation response in Duolingo), such information could significantly improve the performance of our LM-KT model. Other future directions include exploring non-language domains, such as math or logic exercises, and controlling for auxiliary objectives such as question topic.

Finally, designing appropriate user studies to evaluate our method is a complex yet critical next step to determine its suitability in a real-world education setting. Our techniques allows control for individual student difficulty, but it leaves open the question of optimal curriculum design using difficulty-directed question generation.

6 Broader Impact

Online education platforms can increase the accessibility of high quality educational resources for students around the world. Adaptive techniques that allow for more individualized learning strategies can help such technologies be more inclusive for students who make less-common mistakes or have different prior backgrounds (Lee and Brunskill, 2012). However, our method is subject to biases found in the training data, and careful consideration of using safe and appropriate data is crucial in an education context. Moreover, our specific use of pre-trained LMs relies on the significant progress of NLP tools for English language – further research and development of these tools for other languages can help ensure our method benefits a larger population of students.

7 Acknowledgements

References

R. C. Atkinson. 1972. Optimizing the learning of a second-language vocabulary. In *Journal of Experimental Psychology*.

K. C. Bloom and T. J. Shuell. 1981. Effects of massed and distributed practice on the learning and retention of second-language vocabulary. In *The Journal of Educational Research*. Taylor & Francis, Ltd.

A. T. Corbett and J. R. Anderson. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. In *User Modeling and User-Adapted Interaction*. Kluwer Academic Publishers.

Shayan Doroudi and Emma Brunskill. 2019. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics Knowledge*, LAK19, page 335–339, New York, NY, USA. Association for Computing Machinery.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.

RK Hambelton and M Jodoin. 2003. Item response theory: models and features.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez-Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, and Adish Singla. 2018. Teaching multiple concepts to forgetful learners. *CoRR*, abs/1805.08322.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Jung In Lee and Emma Brunskill. 2012. The impact on individualizing student models on necessary practice opportunities. In *EDM*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Frederic M. Lord. 1980. Applications of item response theory to practical testing problems.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. 2020. Progen: Language modeling for protein generation.

Tong Mu, Shuhan Wang, Erik Andersen, and Emma Brunskill. 2018. Combining adaptivity with progression ordering for intelligent tutoring systems. New York, NY, USA. Association for Computing Machinery.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *NeurIPS*, pages 505–513.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks.

D. Rohrer. 2009. The effects of spacing and mixing practice problems. In *Journal for Research in Mathematics Education*. National Council of Teachers of Mathematics.

B. Settles, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani. 2018. Second language acquisition modeling. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. ACL.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Dongmin Shin, Yugeun Shim, Hangyeol Yu, Seewoo Lee, Byungsoo Kim, and Youngduck Choi. 2020. Saint+: Integrating temporal features for ednet correctness prediction.

David Thissen and Robert J Mislevy. 2000. Testing algorithms.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2017. Graph grammar-based controllable generation of puzzles for a learning game about parallel programming. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, FDG '17, New York, NY, USA. Association for Computing Machinery.

David J. Weiss and G. Gage Kingsbury. 1984. Application of computerized adaptive testing to educational problems. In *Journal of Educational Measurement*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A APPENDIX

A.1 Dataset Details

The 2018 Duolingo Shared Task on Second Language Acquisition Modeling (Settles et al., 2018) dataset contains questions and responses for Duolingo users over the first 30 days of learning a second language. The dataset contains three different question types: reverse translate (free response translation of a given prompt in the language they are learning), reverse tap (a selection-based equivalent of reverse translate), and listen, where students listen to a vocal utterance. We focus on the reverse translate question type for English-speaking students learning French and Spanish. The dataset size for French learners (1.2k users) is roughly half the size of that for Spanish learners (2.6k users).

Because the original dataset was intended for per-token error prediction, each question has per-token information that includes whether the student translated the token correctly, as well as Universal Dependencies tags such as part of speech and morphology labels. We use the full question text, rather than individual tokens, for our task, and combine the labels such that if a Duolingo user incorrectly translated one or more tokens in a question, the entire question is marked incorrect. We do not use any additional features.

We use the publicly provided train/dev/test splits from the Shared Task, which are temporally ordered in sequence. We therefore construct student states by tracking user IDs throughout the datasets and appending each new question and response to the current student state. When evaluating our LM-KT model, we use the true responses of preceding questions in the test set to form the student state for a given question. Overall, we find that the dataset is severely imbalanced (as in the original task) - about 30% of questions are answered incorrectly across students studying both French and Spanish.

Finally, we create a held-out set of Duolingo questions for both French and Spanish learners to create the training data for our question generation model. From a set of random student states, we select questions from this set and use a trained LM-KT model to assign the difficulty score. In practice, this held-out set can come from any source, not just Duolingo data.

A.2 Model Training Details

To train both our LM-KT knowledge tracing model and our question generation model, we use the pre-trained OpenAI GPT-2 model from the HuggingFace Transformers library (Wolf et al., 2020). For question generation, we modify the library to add a linear layer and the modified loss function for question generation from Section 3.

We use 1 NVIDIA TitanXP GPU with 12GB of memory available. Because the maximum input sequence length of the GPT-2 model we use is 1024 tokens, we resize all inputs to the last 1024 tokens before training. We report results for an LM-KT model trained for 13k steps with the default batch size of 2 and learning rate of 5e-5, and a Question Generation model trained for 25k steps with the same batch size and learning rate. The total compute time to train both models was 2.5 hours for each language learning task.

A.3 Question Generation Details

For both French and Spanish question generation models, we select 15 students unseen during training and generate 30 questions across 9 difficulties from 0.1 to 0.9, using nucleus sampling (Holtzman et al., 2020) ($p = 0.99$) with a maximum output length of 20 tokens. We also vary a repetition penalty (Keskar et al., 2019) that penalizes for previous tokens (including those in the student state). Lastly, we resize all prompts (student state and target difficulty) to fit into the GPT-2 Model by taking the most recent 1024 tokens, as in training. This is a limitation of our work, as the full student history is not able to be considered for students who have answered a large set of questions.

A.4 Additional Question Generation Outputs

Our question generation model demonstrates the ability to generate novel questions that do not exist in the entire Duolingo question dataset, especially when a sampling penalty is applied to encourage more diverse outputs. However, this comes at a cost to fluency. Below we include a set of outputs generated by our model for 1 Spanish student and 1 French student from the Duolingo dataset, with a target difficulty of $d = 0.1$, and both with and without a repetition penalty. We observe that while applying a penalty results in a far more novel questions generated, several of these are also non-fluent, using a combination of manual judgement and the Python language-check package (https://pypi.org/project/language-check/).

Table 4: Random selection of generated questions for one Spanish learner with for a a target difficulty of $d = 0.1$. *Italic questions* are novel, **bold questions** are judged to be non-fluent.

Spanish (w/ Penalty)	Spanish (No Penalty)
accordingly he does it.	*he mixes coffee with milk.*
clean your room or close!	*the cuts are not big.*
clean your room!	*the gallery is enormous.*
he mixes coffee with water.	*the horses are not natural.*
how many elephants eat cheese or fish?	*the men drink a beer.*
i know about that book.	***they probably do not think me.***
october finds him maximum distance from here today!	*we can desk a book.*
please clean your room!	from september to december
please open your bottle or newspaper?	according to you, it is yellow.
she blames us!	clean the mirror.
she reads us lunchtime newspapers.	i do not know it.
she reads your letters.	i read the newspaper.
those ducks drink water.	i want a sandwich without cheese.
we can abandon him.	june starts tomorrow.
what book have they Chosen me so far?	she reads the calendar.
you can control her water.	the plates are not big.
you can establish two properties.	we are following the clue.
your house is very put- pretty!	we drink quickly.
previously on television	we eat strawberries.
you can create the menu.	you can control the water.
you write letters.	you can create the menu.
your hat is gray	you can establish a restaurant.

Table 5: Random selection of generated questions for one French learner with for a a target difficulty of $d = 0.1$. *Italic questions* are novel, **bold questions** are judged to be non-fluent.

French (w/ Penalty)	French (No Penalty)
do these children have beans?	*do you have three daughters?*
do they come here often?	*do you like this?*
do we come here often or frequently?	*do you speak french?*
do we have chocolate or water?	*where do the children read?*
do we have coffee here or elsewhere?	do you come here often?
do we have coffee together or onsocks	do you want to dance with me?
do we like to walk distance from one-to two?	some apples, which ones?
do we like to walk together or apart?	corridor or window?
do we speak soon or after tomorrow?	neither do we!
is he chinese or Russian?	you are important.
is he chinese or french?	you are important.
is he sleeping or going out time?	are we going to your place or mine?
map ofis suggests an area.	corridor or window?
otherwise if i want to eat vegetables or fish they regionally cheese, it's meat.	do you have a boyfriend?
some apples of your apple.	do you like to walk?
where do we live today?	neither do we!
where does he go after that jacket?	otherwise, i want a child!
where does she go?	the men are calm and rich.
which ones do not fall victim to be sold?	the parties are in august.
beans and bread	we are reading your letters.
corridor or window?	where do we live?
neither do we!	you eat pork and bread