Part II Project Proposal: Modelling Second Language Acquisition

1 Introduction and description of the work

Second Language Acquisition (SLA) is a complex process influenced by a learner's prior knowledge, memory constraints, and exposure to new linguistic forms. Over 1.5 billion people worldwide are learning a second language, often through platforms such as Duolingo. These platforms collect extensive logs of learner behaviour, recording every correct and incorrect response to language exercises, yet can struggle to adapt to individual learners. For example, a student may struggle with certain grammatical constructions, or may repeatedly forget a word after long intervals. Without personalised support, learners can lose motivation and fail to make progress.

The aim of this project is to apply natural language processing (NLP) and machine learning (ML) to build systems to predict when learners are likely to make mistakes. This project will utilise the Duolingo Second Language Acquisition Modelling (SLAM) 2018 shared-task dataset [1], which contains token-level correctness labels for Engish, French, and Spanish learners. Specifically, given the closest reference answer to a Duolingo reverse translation task, the system should predict which tokens a learner will answer correctly and which they will answer incorrectly.

2 Starting Point

From the Tripos, relevant courses include Part IA Machine Learning and Real-World Data, Part IB Formal Models of Language, and Part IB Artificial Intelligence. I will also be taking the Part II NLP module in Michalemas 2025.

The choice of language for this project will be Python, since it has many libraries and tools for NLP-related tasks. I have experience using Python in prior coursework.

This project is inspired by the Duolingo SLAM 2018 shared task, for which I have read the overview paper [2], as well as some of the approaches used for the task, such as the NYU system [3]. In addition to this, I have also come across more recent works involving question generation [4] [5].

3 Substance and Structure of the Project

The SLAM 2018 dataset contains millions of tokens corresponding to answers submitted by over 6,000 students within their first 30 days of using Duolingo, for the 3 different language tracks. For each user, the first 80% of their exercises were placed in the train set, with the next 10% in the dev set, and the final 10% in the test set. The dataset includes user-level identifiers and metadata, exercise-level features, token-level linguistic annotations, and temporal information.

This dataset is publicly available under the CC BY-NC 4.0 license, and approval from the Ethics Committee is required for its use in this project.

The objectives of this project are:

Core task: Re-implement and evaluate the baseline logistic regression model, and then the NYU system, which uses Gradient Boosted Decision Trees (GDBTs) trained on linguistically and cognitively inspired features. The implementation will use LightGBM or an equivalent GDBT framework, with models trained independently for each language track and a combined multilingual model. Following the NYU approach, per-language predictions will be averaged with the multilingual model to produce final predictions.

Extensions:

- Extend the GDBT model by introducing additional features and variations inspired by other SLAM submissions or my own analyses. For example, I can experiment with alternative GDBT frameworks and feature sets.
- Implement a basic knowledge tracing (KT) system, extending to incorporate newer works such as language models (LM) and neural methods, to produce an LM-KT model.
- Explore adaptive question generation based on KT for personalised practice. This would involve building a system to dynamically select practice items of suitable difficulty to extend beyond the prediction task of the shared dataset.
- If time permits, I am also open to re-implementing and extending another submission from the shared task

4 Success criteria and evaluation plan

The project will be considered successful if re-implementations achieve similar performance to the original reported results in the shared task. The primary evaluation metric is AUC, with F1 score also being reported, following the shared task metrics.

I will conduct error analysis for the GDBT system, identifying patterns where the system makes mistakes. I will also perform ablation studies on feature groups to measure their effect.

Similarly, KT and LM-KT will be evaluated using AUC/F1, being considered a success if they lead to performance comparable to the shared task submissions.

The question generation system will be evaluated across multiple dimensions, such as difficulty control, novelty, grammatical correctness/fluency, and latency/efficiency, and can be considered a success if performance is reasonable across these dimensions. The exact metrics used to evaluate will depend on the implementation carried out.

5 Work plan

Michaelmas

Slot 1: 10th October - 23rd October

Set up repository on GitHub and download the SLAM 2018 dataset

Re-implement the official logistic-regression baseline used in the task

Milestone: Baseline system implemented and evaluated against official metrics

Slot 2: 24th October - 6th November

Study the NYU approach in detail

Implement NYU feature groups (exercise, word, user, position, temporal)

Re-parse sentences from the dataset in line with NYU approach

Milestone: Pre-processing and feature engineering complete

Slot 3: 7th November - 20th November

NLP module assignment: 13th November

Train and tune hyper-parameters for the LightGBM (GDBT) system

Finish NYU system implementation

(Extension) Experiment with new feature groups applied to the system

Milestone: Implementation of GDBT system finished

Slot 4: 21st November - 4th December

NLP module assignment: 4th December

Evaluate baseline and NYU system

Perform ablation studies by feature group

Milestone: Core evaluation complete, core success criteria achieved

---

End of Michaelmas / Christmas Break

Slot 5: **5th December - 18th December**

Identify any systematic errors in predictions; update system where feasible

Research Knowledge Tracing, Language Models, and relevant methods for next slot

Work through PyTorch/TensorFlow tutorials

Milestone: Familiarised with concepts and ready to implement KT systems

Slot 6: **19th December - 1st January**

Begin implementing basic Knowledge Tracing system

Milestone: Basic KT system complete

Slot 7: **2nd January - 15th January**

Incorporate contextualised embeddings / language models into KT

Milestone: LM-KT system complete

Lent

Slot 8: **16th January - 29th January**

Draft progress report and slides for presentation

Begin work on question generation system

Milestone: Progress report draft complete, presentation slides prepared

Slot 9: **30th January - 12th February**

Submit progress report

Finish question generation system

Milestone: Progress report submitted; QG system complete

Slot 10: **13th February - 26th February**

Full evaluation of KT, LM-KT, and QG systems

Compare with core implementations

Milestone: Evaluation of all systems complete

Slot 11: 27th February - 12th March

Buffer slot for catching up on any remaining implementation or evaluation tasks

Milestone: All systems finished and evaluated

End of Lent / Easter Break

Slot 12: 13th March - 26th March

Draft introduction and preparation chapters

Milestone: Introduction and preparation chapter draft complete

Slot 13: 27th March - 9th April

Complete full draft of dissertation; submit to project supervisor for feedback

Milestone: Full dissertation draft complete

Slot 14: 10th April - 23rd April

Receive feedback from project supervisor and adjust dissertation

Milestone: Second draft implementing feedback complete

Easter

Slot 15: 24th April - 7th May

Finalise dissertation, implementing any further feedback

Milestone: Further drafts completed incorporating feedback

Slot 16: 8th May - 15th May

Proofread dissertation and only make minor edits

Submit dissertation and source code

Milestone: Completed dissertation and source code submitted

6 Resources

I will be utilising my personal machine with the following specifications:

- 2.1 GHz Intel Core i7-1260P CPU

- Intel Irix Xe integrated GPU
- 16GB RAM, 512GB SSD, Linux

I accept full responsibility for this machine and I have made contingency plans to protect myself against hardware and/or software failure. Code will regularly be backed up on GitHub. The dissertation will be written locally and will be regularly backed up to Google Drive, along with other relevant files.

Should my laptop suddenly fail, I have access to a secondary personal machine, from which I can restore my work from GitHub/Google Drive and continue.

For additional computing requirements, I request access to the department's High-Performance Computing (HPC) facilities. If access is unavailable, I will utilise cloud platforms such as Google Colab or Microsoft Azure, with personal credits used where free versions are not sufficient. I estimate requiring up to 100 hours of machine time.

References

[1] Burr Settles. Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM), 2018.

[2] Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. Second language acquisition modeling. In Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[3] Alexander Rich, Pamela Osborn Popp, David Halpern, Anselm Rothe, and Todd Gureckis. Modeling second-language learning from a psychological perspective. In Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–230, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[4] Megha Srivastava and Noah Goodman. Question generation for adaptive education. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 692–701, Online, August 2021. Association for Computational Linguistics.

[5] Peng Cui and Mrinmaya Sachan. Adaptive and personalized exercise generation for online language learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

10184–10198, Toronto, Canada, July 2023. Association for Computational Linguistics.