SLAM: Deep Second Language Acquisition Modeling

Nathan Dalal, Arjun Manoj, Nidhi Manoj

nathanhd@stanford.edu, arjunmanoj333@gmail.com, nmanoj@stanford.edu

Stanford | ENGINEERING
Computer Science

I. Introduction

- Deep learning techniques enable stronger analysis of tasks such as language acquisition and allow for more personalized online learning
- Second Language Acquisition (SLA) involves learning a target language (L2) from the student's source language (L1)

II. SLAM Dataset

- Duolingo's Second Language Acquisition Modeling (SLAM) dataset is the largest dataset for language acquisition available
- Provides large corpus of student data to trace how users learn a new language through many translation exercises
- Contains exercise information from 6.4K students during the first 30 days of learning a language on Duolingo
- English Track Dataset Size
  - Train: 824K exercises, 2.6M tokens
  - Validation: 115K exercises, 387K tokens

III. Prediction Task

- Task was released as a public, worldwide challenge.
- For each word (token), we would like to predict whether the student got it correct or incorrect. This is a binary classification task
- Input can be passed into models at various levels. We propose to apply various recurrent architectures to SLAM, at the following levels:
  - instance level (one word at a time)
    - Model: Logistic Regression
  - exercise level (exercise's sequence of words)
    - Model: ExerciseLSTM
  - user level (one student at a time)
    - Model: UserLSTM

learner:	wen	can		help	
reference:	when	can	I	help	?
label:	✗	✓	✗	✓	

Figure 1. We see a typo on 'when' and a missing pronoun 'I' in the input phrase above.

IV. Evaluation Metric

- Evaluate using primary metric AUC and secondary metric F1

V. Approach

Figure 2. Example of student learning French from English on Duolingo.

```
# prompt:The bee is an insect.
# user:Nsr+jY0A  countries:US  days:5.496  client:ios  session:practice  format:reverse_translate  time:24
+77qRODw0501  L'          DET    Definite=Def|Gender=Fem|Number=Sing|fPOS=DET++                    det    2  1
+77qRODw0502  abeille     NOUN   Gender=Fem|Number=Sing|fPOS=NOUN++                                nsubj  5  1
+77qRODw0503  est         VERB   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++  cop    5  1
+77qRODw0504  un          DET    Definite=Ind|Gender=Masc|Number=Sing|PronType=Dem|fPOS=DET++      det    5  0
+77qRODw0505  insecte     NOUN   Gender=Masc|Number=Sing|fPOS=NOUN++                               ROOT   0  1
```

Figure 3. Training data representation of a student exercise, as presented in Figure 2.

Exercise Level Features

See first two lines with hash (#) in Figure 2.

- Prompt
- User
- Country
- Days
- Client (web, ios, android)
- Session (lesson, practice, test)
- Format (reverse translate, reverse tap, listen)
- Time to submit exercise

Word Level Features

Any of the bottom five lines in Figure 3.

- Token encoded as a MUSE embedding (size 300)
- Part of Speech
- Dependency Label
- Dependency Head Edge

VI. Models

Experiment 1: v1 Logistic Regression
- Logistic Regression without user features
- Included MUSE word embeddings

Experiment 2: v2 Logistic Regression
- Now with user features which improved model from baseline

Experiment 3: Exercise LSTM (shown on left)
- Model an exercise as a sequence of words
- LSTM over word phrases
- Build deeper feature encodings & stacked model

Experiment 4: Generative Modeling
- Used a conditional variational autoencoder (CVAE) to synthesize more instances

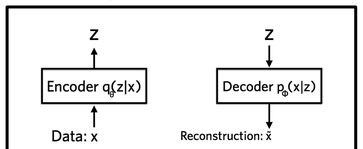

Figure 5. Conditional variational autoencoder

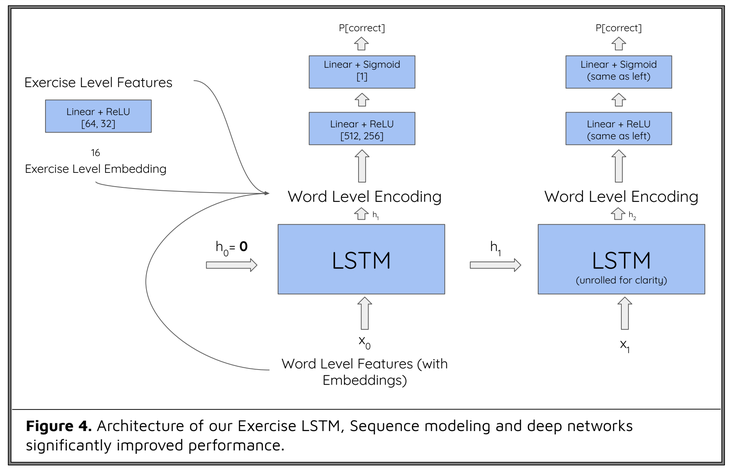

Figure 4. Architecture of our Exercise LSTM, Sequence modeling and deep networks significantly improved performance.

VII. Results

Model (employ real data unless specified)	AUROC	F1	Accuracy
State-of-the-art Sana Labs	**0.861**	0.561	----
***Our Exercise LSTM	**0.840**	0.423	0.880
***Our User Logistic Regression	0.781	0.203	0.863
Duolingo Baseline	0.774	0.190	----
***Our Non-User Logistic Regression	0.730	0.111	0.858
***Our Logistic Regression (synthetic + real)	0.744	0.236	0.861
***Our Logistic Regression (synthetic)	0.591	0.269	0.672

Table 1. Performance on English track. Our model results on the validation set (*) compared to Duolingo baseline and the state-of-the-art on the test set. Exercise LSTM is good for 7th place on the English SLAM leaderboard and comparable to 4th - 6th place. Synthetic data must be improved and integrated into sequence models.

VIII. Experiments and Analysis

- Xavier initialization, Adam optimizer, learning rate reduced on plateau started at 0.001, and weight decay set to 0.00001, for 50 epochs
- Test and validation set performance are consistent, indicating that model generalizes to unseen data well

AUROC on validation set			
	English	French	Spanish
Exercise LSTM	0.840	0.797	0.775
Deeper LSTM	0.837	0.798	0.776
Stacked LSTM	0.821	0.780	0.765

AUROC on test set			
	English	French	Spanish
Exercise LSTM	0.835	**0.795**	0.776
Deeper LSTM	**0.837**	0.793	**0.778**

- User encoding and MUSE word embeddings are important
- Strong balance of predicting positive and negative exercises, despite class imbalance
- Deeper models improve performance, because of avoidable bias

Addressing Data Bottleneck through Generative Synthesis
- We ran the CVAE first (1) without labels and then (2) with labels learning both the correct/incorrect prediction distributions at the same time. Finally (3) we ran two separate CVAE models to learn and sample the correct and incorrect (labels 0/1) prediction distributions
- We then sample from our CVAE distributions to form synthetic data, which we use by sending back into our training models.
- Accurate learning data synthesis needs more work.

VIII. Acknowledgements

Thanks to Chris Piech (mentorship), Duolingo (GPUs), and Exxact (GPUs).

Selected References

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

[2] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.