SLAM: Deep Second Language Acquisition Modeling

Nathan Dalal, Nidhi Manoj

{nathanhd, 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)
- We propose to apply various recurrent architectures to SLAM, at the instance, exercise, and user levels.

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

- Released as a public, worldwide challenge. We focus on the English language track.
- 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 tried a model for each of these 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. Approach

Figure 2. Example of a 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

V. 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

Experiment 4: User LSTM
- Model a user as a group of exercises
- Allows modeling of forgetting over time
- Sequence of ExerciseLSTM (shown on left) passing the previous exercise embeddings
- Did not optimize correctly, results omitted

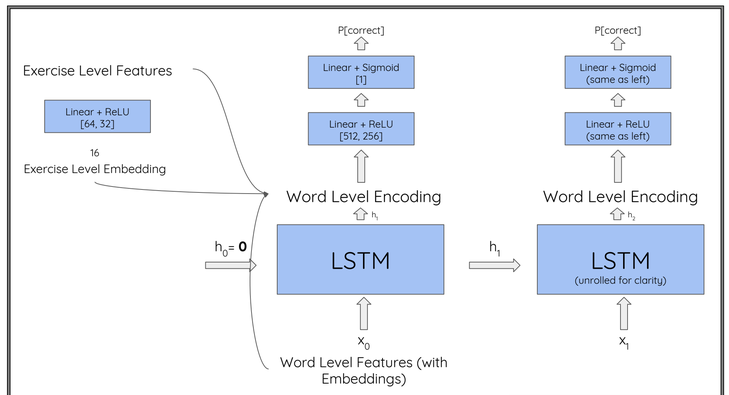

Figure 4. Architecture of our Exercise LSTM, Sequence modeling and deep networks significantly improved performance.

VI. Results

Model	AUC	F1	Accuracy
State-of-the-art Sana Labs	**0.861**	0.561	–
∗Our ExerciseLSTM	**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

Figure 5. 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.

VI. Evaluation Metric

- Evaluate using primary metric AUC and secondary metric F1

VII. Experiments and Analysis

- Xavier initialization, Adam optimizer, learning rate reduced on plateau, and weight decay set to 0.00001
- Logistic Regression with user embedding takes ~6 minute/epoch to train. LSTM takes ~7 minutes/epoch to train
- Tuned models for various initial learning rates and batch sizes

Confusion Matrix for Exercise LSTM		
	Actual Positive	Actual Negative
Predicted Positive	TP = 16999	FP = 8405
Predicted Negative	FN = 38359	TN = 323611
	Total = 55358	Total = 332016

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

VII. Future Work

- Fix UserLSTM and try other ways of encoding user information
- Train deeper models, tune hyperparameters, dropout and regularization
- Train on more languages!
- Beat the state of the art!

VIII. Acknowledgements

- Thanks to Michael Hahn, our CS224N mentor for his guidance. Thanks to Arjun Manoj, an external contributor on the project.

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.