Modeling Second-Language Learning from a Psychological Perspective

Alexander Rich, Pamela Osborn Popp, David Halpern, Anselm Rothe, & Todd Gureckis

Computation & Cognition Lab

NYU

Intro

Second Language Acquisition Modeling

The Duolingo SLAM competition provided log data from thousands of users, posing the challenge of predicting patterns of future translation mistakes in held-out data.

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

Hypothesis Insights from the large literature on the psychology of learning, memory, and motivation might improve machine learning model predictions.

Data set

Data were provided from three tasks from the first 30 days of user sessions:

- free reverse translate
- reverse translate from word bank
- transcribe an audio clip

		TRAIN	DEV	TEST
Track	Users	Tokens (Err)	Tokens (Err)	Tokens (Err)
English	2.6k	2.6M (13%)	387k (14%)	387k (15%)
Spanish	2.6k	2.0M (14%)	289k (16%)	282k (16%)
French	1.2k	927k (16%)	138k (18%)	136k (18%)
Overall	6.4k	5.5M (14%)	814k (15%)	804k (16%)

Three language tracks:

English learners (who speak Spanish)
Spanish learners (who speak English)
French learners (who speak English)

TRAIN: first 80% of exercises
DEV: next 10% of exercises
TEST: last 10% of exercises

Model predicted held-out last tenth for the same users from train and dev.

Example of exercise data format:

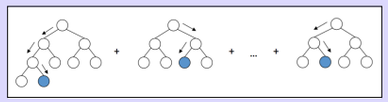
```
# user:D2inSf5+  countries:MX  days:2.689  client:web  session:practice  format:reverse_translate  time:6
oMGsnnH/0101  When  ADV   PronType=Int|fPOS=ADV++WRB                          advmod  4  1
oMGsnnH/0102  can   AUX   VerbForm=Fin|fPOS=AUX++MD                          aux     4  0
oMGsnnH/0103  I     PRON  Case=Nom|Number=Sing|Person=1|PronType=Prs|fPOS=PRON++PRP  nsubj   4  1
oMGsnnH/0104  help  VERB  VerbForm=Inf|fPOS=VERB++VB                         ROOT    0  0
```

Supported by NSF grants DRL-1631436 and BCS-1255538

Modeling

Architecture

Gradient Boosting Decision Trees (GBDT)



GBDT can extract complex interactions among features, but is faster to train than deep learning and more easily integrates diverse inputs.

Implemented with the `LightGBM` library in python.

Feature engineering

Categorical variables were either one-hot encoded or handled natively by `LightGBM`.

Exercise features *exercise number*, *client*, *session*, *task format, time to complete exercise, days since user started language on Duolingo*

Word features *original word token* (ID), *root word token* (via spaCy), *part of speech*, *morphological features* (via Google SyntaxNet), *dependency edge label*, *word length*

based on external sources

- *word frequency* (in natural language)
- *age of acquisition* (age at which English native speaker children exhibit the English word in their vocabulary)
- *Levenshtein edit distance* (between word token and translation from Google Translate, scaled by length of longer word)

Cognate
🍅 = tomato = tomate = 🍅

False friend
👜 = bag ≠ bague = 💍

User features *user ID, "motivation"* (average number of exercises within burst [bursts were separated by at least 1 hour], total number of bursts), *"diligence"* (variability in exercise start times)

Positional features *tokens of previous & next word, including their part of speech*

Temporal features *number of times exercise repeated*, *past performance on each word* (tracked with four different temporal decay rates)

Results

Model performance

Predictive accuracy was measured as area under the ROC curve (AUROC).

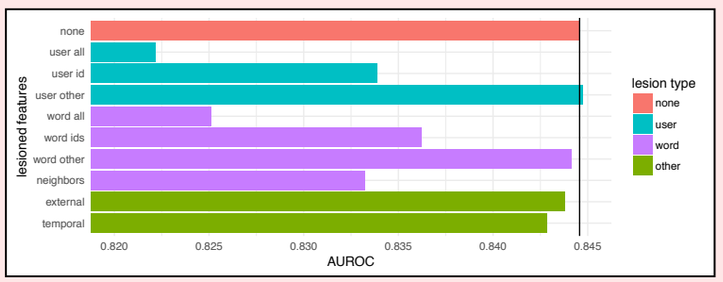
Our GBDT model scored within .01 of the winner's AUROC for each track.

AUROC scores

	baseline	NYU	winner
English	0.774	**0.859**	0.861
Spanish	0.771	**0.854**	0.857
French	0.746	**0.835**	0.838

Feature removal analysis

We systematically lesioned and retrained the model to evaluate which groups of features were most important to the model's predictive accuracy (AUROC).



User ID and word ID accounted for much of our model's predictive power.

In their absence, the other features were useful.

The temporal features added less benefit than we expected.

The external word features (frequency, age of acquisition, Levenshtein distance) lead to some improvement.

Discussion

We created a number of *features* inspired by concepts in psychology but our *prediction engine* came from machine learning and did not specifically model the *process* of learning.

How can we adapt psychological *theories* of learning for such modeling?

How useful are concepts and theories from psychology for making predictions for *individual* users (instead of *group level* predictions)?