Second Language Acquisition Modeling

2018 NAACL / BEA Shared Task Report

Burr Settles
Duolingo

Chris Brust
Duolingo

Erin Gustafson
Duolingo

Masato Hagiwara
Duolingo

Nitin Madnani
Educational Testing Service
why should we care about modeling second language acquisition?
people learning a second language

1,200,000,000
(~16% of the world’s population)

~800M satisfy three properties:
- learning English
- in a developing country
- to gain more opportunity

(Source: British Council)
86% mobile device access

64% toilet access

enormous potential for computer-based, adaptive language-learning!
SLA Modeling

accurately model **what** language-learners know and **how well** they know it ...
SLA Modeling

... and do so in a **personalized** way
(that adapts + learns over time)
Learner Modeling in Other Domains

DataShop (Koedinger et al., 2010)

- **150 public** research data sets
- mostly **math + physics** domains, largely multiple-choice items
- still relatively **small**:
  - 71 avg students (5k max)
  - 880 avg instances (1.5M max)
Our Goals for the SLAM Task

• facilitate dialog among ML/NLP/CogSci fields through a common large-scale empirical task

• accessible, familiar data format + task definition (e.g., classification similar to other shared tasks)

• include languages other than English

• start with beginners who are learning over time
launched in 2012 (CMU research spinoff)
more than 200 million students globally
currently 79 courses (incl. Irish, Esperanto, + Klingon!)
expanding to 93 courses (incl. Arabic + Hindi!)
content is FREE

GOOGLE  Best of the Best
APPLE  App of the Year
TECHCRUNCH  Education Startup of the Year
The Data

reverse_translate

reverse_tap

listen
The Data

**prompt:** cuándo puedo ayudar

when (can i|am i able to) (help (out||)assist)

**student:** wen can help

**reference:** when can I help

**label:** 1 0 1 0
The Data

<table>
<thead>
<tr>
<th>IDs</th>
<th>reference answer tokens</th>
<th>morpho-syntactic features</th>
<th>labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>oMGsnnH/0101</td>
<td>When</td>
<td>ADV PronType=Int</td>
<td>fPOS=ADV++WRB</td>
</tr>
<tr>
<td>oMGsnnH/0102</td>
<td>can</td>
<td>AUX VerbForm=Fin</td>
<td>fPOS=AUX++MD</td>
</tr>
<tr>
<td>oMGsnnH/0103</td>
<td>I</td>
<td>PRON Case=Nom</td>
<td>Number=Sing</td>
</tr>
<tr>
<td>oMGsnnH/0104</td>
<td>help</td>
<td>VERB VerbForm=Inf</td>
<td>fPOS=VERB++VB</td>
</tr>
</tbody>
</table>
The Data

user + session-level metadata

<table>
<thead>
<tr>
<th>user:XEinXf5+</th>
<th>countries:CO</th>
<th>days:2.678</th>
<th>client:web</th>
<th>session:practice</th>
<th>format:reverse_translate</th>
<th>time:6</th>
</tr>
</thead>
<tbody>
<tr>
<td>oMGsnnH/0101</td>
<td>when</td>
<td>ADV</td>
<td>PronType=Int</td>
<td>fPOS=ADV++WRB</td>
<td>advmod</td>
<td>4</td>
</tr>
<tr>
<td>oMGsnnH/0102</td>
<td>can</td>
<td>AUX</td>
<td>VerbForm=Fin</td>
<td>fPOS=AUX++MD</td>
<td>aux</td>
<td>4</td>
</tr>
<tr>
<td>oMGsnnH/0103</td>
<td>I</td>
<td>PRON</td>
<td>Case=Nom</td>
<td>Number=Sing</td>
<td>Person=1</td>
<td>PronType=Prs</td>
</tr>
<tr>
<td>oMGsnnH/0104</td>
<td>help</td>
<td>VERB</td>
<td>VerbForm=Inf</td>
<td>fPOS=VERB++VB</td>
<td>ROOT</td>
<td>0</td>
</tr>
</tbody>
</table>
The Data

oMGsnnH/0101 When ADV PronType=Int|fPOS=ADV++WBRB 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

W+QU2fm70301 He PRON Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP nsubj 3 0
W+QU2fm70302 's AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=AUX++VBZ aux 3 1
W+QU2fm70303 wearing VERB Tense=Pres|VerbForm=Part|fPOS=VERB++VBG ROOT 0 0
W+QU2fm70304 two NUM NumType=Card|fPOS=NUM++CD nummod 5 0
W+QU2fm70305 shirts NOUN Number=Plur|fPOS=NOUN++NNS dobj 3 0

vOeGrMgP0101 We PRON Case=Nom|Number=Plur|Person=1|PronType=Prs|fPOS=PRON++PRP nsubj 2 0
vOeGrMgP0102 eat VERB Mood=Ind|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBP ROOT 0 1
vOeGrMgP0103 cheese NOUN Degree=Pos|fPOS=ADJ++JJ dobj 2 1
vOeGrMgP0104 and CONJ fPOS=CONJ++CC cc 2 0
vOeGrMgP0105 they PRON Case=Nom|Number=Plur|Person=3|PronType=Prs|fPOS=PRON++PRP nsubj 6 0
vOeGrMgP0106 eat VERB Mood=Ind|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBP conj 2 1
vOeGrMgP0107 fish NOUN fPOS=X++FW dobj 6 0
Data Partitions (Sequential)

80% TRAIN  10% DEV  10% TEST

← USERS →

TIME (30-day sampling window) →
Duolingo’s three largest courses (~1/3 of users)

<table>
<thead>
<tr>
<th>Language Track</th>
<th>USERS</th>
<th>TRAIN (tokens)</th>
<th>DEV (tokens)</th>
<th>TEST (tokens)</th>
<th>TOTAL (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (EN ← ES)</td>
<td>2,593</td>
<td>2,622,958</td>
<td>387,374</td>
<td>386,604</td>
<td>3,396,936</td>
</tr>
<tr>
<td>Spanish (ES ← EN)</td>
<td>2,643</td>
<td>1,973,558</td>
<td>288,864</td>
<td>282,181</td>
<td>2,544,603</td>
</tr>
<tr>
<td>French (FR ← EN)</td>
<td>1,213</td>
<td>926,657</td>
<td>137,571</td>
<td>135,525</td>
<td>1,199,753</td>
</tr>
<tr>
<td>TOTAL (All 3 Tracks)</td>
<td>6,449</td>
<td>5,523,173</td>
<td>813,809</td>
<td>804,310</td>
<td>7,141,292</td>
</tr>
</tbody>
</table>
Other Details

- **evaluation**: AUC (official metric) + F1
- **development phase (TRAIN + DEV)**: 8 weeks
- **test phase (TEST)**: 10 days
  - blind TEST set submissions via CodaLab
  - teams allowed to use both TRAIN+DEV to train
Participation Pipeline

- TRAIN/DEV downloads: 245
- TEST downloads: 102
- mailing list signups: 78
- CodaLab signups: 26
- CodaLab submissions: 15
- paper submissions: 11
Official Results
Official Results

- English
- Spanish
- French

- Linear models
Official Results

Tree Ensembles (GBDT, RF)

Linear models
Official Results

English

French

Spanish

NYU*
TMU*
CECL*
Camb r idge
UCSD
LambdaLab
nihaln a
Grotoco
jilljenn
ymatus e
renhk
SLAM_baseline
SanaLabs
singsound

• Linear models
• Tree Ensembles (GBDT, RF)
△ RNN (across exercises)
Official Results

English

Spanish

French

- Linear models
- Tree Ensembles (GBDT, RF)
- RNN (across exercises)
- Hybrid (RNN+GBDT)
Does the Algorithm Matter?

<table>
<thead>
<tr>
<th>Fixed effects (algorithm choices)</th>
<th>Effect</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.786</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Recurrent neural network</td>
<td>+ .028</td>
<td>.012</td>
</tr>
<tr>
<td>Decision tree ensemble</td>
<td>+ .018</td>
<td>.055</td>
</tr>
<tr>
<td>Linear model (e.g., IRT)</td>
<td>− .006</td>
<td>.541</td>
</tr>
<tr>
<td>Multitask model</td>
<td>+ .023</td>
<td>.017</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effects</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>User ID</td>
<td>±.086</td>
</tr>
<tr>
<td>Team ID</td>
<td>±.013</td>
</tr>
<tr>
<td>Track ID</td>
<td>±.011</td>
</tr>
</tbody>
</table>

linear mixed-effects analysis of learning algorithms
Example Multitask Approaches

NYU (Rich et al., 2018) — 3rd

TMU (Kaneko et al., 2018) — 4th
Other Algorithm Notes

• **linear classifiers** are effectively **item response theory** models, specifically AFMs (Cen et al., 2008)

• the **RNN systems** are examples of **deep knowledge tracing** (Piech et al., 2015), an extension of BKT

• the only linear model to rank in the top 5 was CECL, which used logistic regression with **feature conjunctions**
  
  • effectively modifies the decision surface to be **nonlinear**

  • RNN **hidden nodes** + GBDT **constituent trees** may be representing these same conjunctions
Does the Feature Set Matter?

<table>
<thead>
<tr>
<th>Features used</th>
<th>Popularity</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word (surface form)</td>
<td></td>
<td>+.005</td>
</tr>
<tr>
<td>User ID</td>
<td></td>
<td>+.014</td>
</tr>
<tr>
<td>Part of speech</td>
<td></td>
<td>−.008</td>
</tr>
<tr>
<td>Dependency labels</td>
<td></td>
<td>−.011</td>
</tr>
<tr>
<td>Morphology features</td>
<td></td>
<td>−.021</td>
</tr>
<tr>
<td>Response time</td>
<td></td>
<td>+.028  *</td>
</tr>
<tr>
<td>Days in course</td>
<td></td>
<td>+.023  .</td>
</tr>
<tr>
<td>Client</td>
<td></td>
<td>+.005</td>
</tr>
<tr>
<td>Countries</td>
<td></td>
<td>+.012</td>
</tr>
<tr>
<td>Dependency edges</td>
<td></td>
<td>−.000</td>
</tr>
<tr>
<td>Session</td>
<td></td>
<td>+.014</td>
</tr>
</tbody>
</table>

Linear mixed-effects analysis of provided features

Time-related features appear to help somewhat.

Morpho-syntactic features seem to hurt slightly.

---
Parsing (+ Alignment) Errors

Along with the tokens themselves we encoded each instance word’s part of speech, morphological features, and dependency edge label. We noticed that some words in the original dataset were paired with the wrong morphological features, particularly near where punctuation had been removed from the sentence. To fix this, we reprocessed the data using Google SyntaxNet.

Cambridge (Yuan, 2018)

NYU (Rich et al., 2018)
**Does the Feature Set Matter?**

A linear mixed-effects analysis of novel features reveals:

<table>
<thead>
<tr>
<th>Features used</th>
<th>Popularity</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word corpus frequency</td>
<td></td>
<td>+.008</td>
</tr>
<tr>
<td>Spaced repetition features</td>
<td></td>
<td>+.013</td>
</tr>
<tr>
<td>L1-L2 cognates</td>
<td></td>
<td>+.001</td>
</tr>
<tr>
<td>Word embeddings</td>
<td></td>
<td>+.020</td>
</tr>
<tr>
<td>Word stem/root/lemma</td>
<td></td>
<td>+.007</td>
</tr>
</tbody>
</table>

> 30 days might make these more useful

More linguistically diverse data might make these more useful.
Can An Ensemble Do Better?
Can An Ensemble Do Better?

stacking weights (across all 3 tracks)

- 1st (RNN+GBDT)
- 2nd (RNN)
- 3rd (GBDT)
Summary

• first SLA modeling task: attracted 15 teams from diverse fields

• learning algorithm choices (RNNs, GBDTs, multitask) appear to be more impactful than clever feature engineering

• morpho-syntactic features did not seem to help, possibly due to systematic parsing (+ alignment) errors

• a more longitudinal SLA modeling task (> 30 days) + more linguistic diversity (multiple L1s; intermediate-advanced) might let psychologically-inspired features be more useful
Questions?

corpus, papers, starter code, etc. available at:
http://sharedtask.duolingo.com

special thanks to: Bozena Pajak, Joseph Rollinson, Hideki Shima, Eleanor Avrunin, Natalie Glance, Anastassia Loukina, Kristen K. Reyher, the BEA workshop organizers, + of course the participating teams!