Optimizing Human Language Learning

Masato Hagiwara
Software Engineer / Researcher

duolingo

Optimizing Human Learning  6/12/2018 (Montréal, Canada)
http://masatohagiwara.net/its2018.html
279. (a) Tra lua korta staras nia domo. Sia kaj flosas tra la rivero. Il estis tra la maro. Li vagas tra montoj kaj maro.  
(b) Across the field is a rose-garden. Stand and wait on the other side of the street. Scotland is there, across the valley.

KE

280. KE=that, used as a link-word joining two sentences (a "conjunction"). Mi kredas, ke vi estas prava, I believe (that) you are right. Here ke joins the sentences Mi kredas and vi estas prava. Mi scias, ke li logas tic, I know (that) he lives there. Mi bedaŭras, ke vi malsanas; I am sorry (regret) (that) you are ill.

In English that is often omitted, but you must not omit the ke. Place a comma before ke, to divide the sentences.

281. Mi čieli ĝis, ke mi estis Esperantistino. Cu vi sentas, ke mi eratas? Ni opinias estas (mi opinias) (It is our opinion, we think), ke tiu estas ido ido. Kial vi supozas, ke tiu ne estas ido? Si pretendas (claim), ke ti estis reĝino. Mi divenas (quarrel), ke vi estas objekto.

Mi respektas lin tial, ke (because, for the reason that) li estas tre modesta. (Tial ke=a strong form of ĉar).

282. Tri gradoj de mar-malsaneco. (1) Mi kredas, ke mi mortas. (2) Mi scias, ke mi ne mortas. " Cu vi opinias, ke mi estis ido?" "Tute ne! Tamen eble mi eraras."

283. Make sentences using ke after some of the words bedaŭro (dishearten, grieve, lombok, comprišo, konfesio, komposado, lego, memorigi, opinio, predikado, pruvado, repondo, ripeto, scienca, taksado, telefaco). Do not forget the comma.  
(b) Complete the following sentences. Ne furgo, ke... Memori, ke... Ne respondu, ke... Kredas tiu, ke... Li ofte ariĝas, ke... Kial vi supozas, ke... Cu vi do tīnas, ke...? Cu vi vere suspektas, ke...

THAT

284. This English word has many meanings, and must be translated accordingly. He says about that, that that "that" that is in your essay is wrong, Li diras pri tio, ke tiu "that", kiu estas en via esso, estus malsaj.  
285. She says (that) he is rich, but for some reason I do not suppose (that) that is true. Tell him (that) I am not a man of that-kind. I feel (that) you are wrong on that matter. She aŭtoras (that) she is right about that question. Do not believe that all that is in that book is true. That proves the earth is round.

KE JES, KE NE

286. The translation of ke jes, ke ne, depends on the context. Ĉu li estas tie? (Mi kredas, ke jes, I believe he is, I think so. Mi supozas, ke ne, I suppose he isn't, I suppose not). Ĉu vi komprenas? (Mi timas, ke ne, I'm afraid not. I fear I don't. Mi scias, ke jes, I know I do).

287. Write similar questions and answers, using ke jes, ke ne, after some of the verbs averti, bedaŭri, diri, esperi, ĝi, insulti, kompreni, kredadi, opinii, ripeti, sciri, skribi, superigeti, suspekti, telefonigi, vidadi (Compare 104-5).

NOTE ON PREPOSITIONS

288. Al, apud, de, dum, en, inter, kun, por, pri, sen, sub, super, sur, tra, trans, are prepositions (96).

289. In English conversation we often end a sentence with a preposition, thus: "What are you sitting on?" "Whom are you thinking of?" "In spite of the famous rule "A preposition is a word you must not end a sentence with"! But in Esperanto you may NOT do this. Write, therefore: "Sur kio vi sidas?" "Pri kio vi pensas?"

290. Translate these questions and answer them. What are you complaining of! Which town do you come from? What are you reading through? What are we fighting for? What is your money in? Whom are you sitting near? What are you doubtful about? Whom are you waiting for? What does an apple grow on? Whom are you speaking to? What are you talking about? Whom are you dreaming of?

"SUBJECT" AND "OBJECT"

291. Most sentences contain (1) a SUBJECT (the thing or person we are talking about) and (2) a VERB (which shows what the subject is or does). Thus, in La suno brilas, and Li estus la subject of the verb brilas, and Li is the subject of the verb estus. The answer to What? Who? asked before a verb, is the subject of the sentence. What shines? (La suno). Who stands? (Li).  
292. Usually the subject comes before the verb. But it need not do so (so, 23). One may say, for example, Brilas sur la suno. *Tie stari li.

*Even "What was the man who asked what you chose that book to read out of for up to?"!
“Learning Esperanto used to be a solitary quest. You could practise it by sitting for weeks with a book and a dictionary, figuring out the rules and memorising the words. But there was usually no professor to correct your mistakes or polish your pronunciation.”

— Jose Luis Penarredonda, BBC, 2018
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)
Most popular language studied on Duolingo in each country
Human
“User attention is expensive”
3,466 users who stayed

3,426 users who quit
Figure 1: # of days since last session

Figure 2: # of sessions (last weeks)

Figure 4: Relative stdev of daily # of sessions
Dirichlet Process Gaussian Mixture Model (DPGMM)

Visualization by t-SNE

Weekdays
- User A
- User B
- User C
- User D

Weekend
Figure 3

Abilities / Latent Trait

- Weekenders
- Nine-to-fivers
- Daily
- Daily, at bedtime
~30%+ retention
2x relearning

Source: (Mezza et al. 2016)
Language
Duolingo Second Language Acquisition Modeling (SLAM) Data

- Language learning
- Multilingual
- Longitudinal
correct: She is my mother and he is my father.

student: she is mader and he is fhader.

label: O O O X X O O O O O O X X
She is my mother and he is my father.

When I can help, I will.
<table>
<thead>
<tr>
<th>Language</th>
<th>Users</th>
<th>TRAIN</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>2.6k</td>
<td>2.6M</td>
<td>387k</td>
<td>387k</td>
</tr>
<tr>
<td>Spanish</td>
<td>2.6k</td>
<td>2.0M</td>
<td>289k</td>
<td>289k</td>
</tr>
<tr>
<td>French</td>
<td>1.2k</td>
<td>927k</td>
<td>138k</td>
<td>136k</td>
</tr>
<tr>
<td>Overall</td>
<td>6.4k</td>
<td>5.5M</td>
<td>814k</td>
<td>804k</td>
</tr>
</tbody>
</table>
2018 Duolingo SLAM Shared Task

- Phrase 1 (8 weeks): TRAIN and DEV sets released
- Phase 2 (10 days): TEST set released
- Metrics: AUROC (Area Under ROC curve) + F1
- 15 teams participated (11 submitted paper)
<table>
<thead>
<tr>
<th>Team</th>
<th>RNN</th>
<th>Ensemble</th>
<th>Multitask</th>
<th>English</th>
<th>Spanish</th>
<th>French</th>
<th>Average</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>SanaLabs</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.861</td>
<td>0.838</td>
<td>0.857</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>singsound</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.861</td>
<td>0.835</td>
<td>0.854</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>NYU</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.859</td>
<td>0.835</td>
<td>0.854</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>TMU</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>0.848</td>
<td>0.824</td>
<td>0.839</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>CECL</td>
<td></td>
<td></td>
<td>✓</td>
<td>0.846</td>
<td>0.818</td>
<td>0.843</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>Cambridge</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.841</td>
<td>0.807</td>
<td>0.835</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>UCSD</td>
<td></td>
<td></td>
<td>✓</td>
<td>0.829</td>
<td>0.803</td>
<td>0.823</td>
<td>7.0</td>
<td></td>
</tr>
<tr>
<td>LambdaLab</td>
<td></td>
<td></td>
<td>✓</td>
<td>0.821</td>
<td>0.801</td>
<td>0.815</td>
<td>7.6</td>
<td></td>
</tr>
<tr>
<td>Grotoco</td>
<td></td>
<td></td>
<td></td>
<td>0.817</td>
<td>0.791</td>
<td>0.813</td>
<td>9.0</td>
<td></td>
</tr>
<tr>
<td>nihalnayak</td>
<td></td>
<td></td>
<td></td>
<td>0.821</td>
<td>0.790</td>
<td>0.811</td>
<td>9.0</td>
<td></td>
</tr>
<tr>
<td>jilljenn</td>
<td></td>
<td></td>
<td></td>
<td>0.815</td>
<td>0.788</td>
<td>0.809</td>
<td>10.7</td>
<td></td>
</tr>
<tr>
<td>SLAM_basline</td>
<td></td>
<td></td>
<td></td>
<td>0.774</td>
<td>0.746</td>
<td>0.771</td>
<td>14.7</td>
<td></td>
</tr>
</tbody>
</table>
#1. SanaLabs
(Osika et al., 2018)

- Combination of GBDT (Gradient Boosted Decision Tree) and RNN (LSTM)
  - GBDT — good for tabular data
  - RNN — good for sequential data
- Additional features (e.g., context, times seen)
#2. singsound
(Xu et al., 2018)

```
0 1 0 0
```

user format

```
She is my mother
```

Context Encoder (bi-LSTM)

Linguistic Feature Encoder (LSTM)

Logistic
Shared Task Takeaways

- Choice of learning algorithm > feature engineering
  - Sequence models (RNN)
  - Tree ensembles (GBDT)
  - Multitask learning
- Mostly beginners
  - vs more longitudinal, diverse user base
Learning
The 2 Sigma Problem

Source: (Bloom 1984)
A Tale of Two Bars
(or, Duolingo’s answers to the two sigma problem)
Le chat et la souris

Translation:
The cat and the mouse
Source: (Streeter 2015) - Users learning es<-en on Duolingo, “una” on listen items
Bayesian Knowledge Tracing

Not Learned

Initial Learning
\( p(L_0) \)

Learned

\( p(T) \)

Acquisition

Guess
\( p(G) \)

Slip
\( p(S) \)
Source: (Streeter 2015)
Item Response Theory (1PL)
\[ \theta_i + \beta_j \]

Additive Factor Model
[Cen et al. 2006]
\[ \theta_i + \beta_j + \gamma_j T_{ijt} \]

Performance Factor Model
[Pavlik et al. 2009]
\[ \theta_i + \beta_j + \alpha_j S_{ijt} + \rho_j F_{ijt} \]

\[ \sum_{k=1}^{K} p^k \prod_{t=1}^{T} \mathcal{B}(\phi_{i,j,t} \cdot \beta^k, v_t) \]

\( i \) student
\( j \) problem/KC
\( T_{ijt} \) number of practice opportunities
\( S_{ijt} \) number of successes
\( F_{ijt} \) number of failures
<table>
<thead>
<tr>
<th>Method</th>
<th>Training log loss</th>
<th>Test log loss</th>
<th>Training AUC loss</th>
<th>Test AUC loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Tracing</td>
<td>0.3429</td>
<td>0.3441</td>
<td>0.3406</td>
<td>0.3460</td>
</tr>
<tr>
<td>Performance Factors Analysis</td>
<td>0.3248</td>
<td>0.3285</td>
<td>0.2774</td>
<td>0.2865</td>
</tr>
<tr>
<td>Additive Factor Model</td>
<td>0.2869</td>
<td>0.3250</td>
<td>0.1629</td>
<td>0.2789</td>
</tr>
<tr>
<td>AFM Mixture (3 components)</td>
<td><strong>0.2818</strong></td>
<td><strong>0.3220</strong></td>
<td><strong>0.1598</strong></td>
<td><strong>0.2760</strong></td>
</tr>
</tbody>
</table>
Forgetting Curve

the probability $p$ of a correct answer as a function of:

- **time** $\Delta$ since it last practice
- **halflife** $h$ in user’s memory

$$p = 2^{-\Delta/h}$$

(Ebbinghaus, 1885; Pimsleur, 1967; Leitner, 1972)
The Leitner System

(Leitner, 1972)

\[ h = 2^x \oplus -x \oplus \]

\[ h = 2^{\Theta} \cdot x \]
Half-life Regression (HLR)
Results: User Experiment

(6 weeks; 3.6M subjects)

all students who returned the next day:

+1.7% (p<0.001)

(ca. April 2013)
“Thanks to the help of Duolingo, I was able to get a job teaching English at the top security academy in my city.”

- Edilson García Vargas
Thanks!

http://masatohagiwara.net/its2018.html

... and we’re hiring!