Second Language Acquisition Modelling

Anton Osika (Sana Labs)
at
Swedish Institute of Computer Science

2018-05-17

2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)

This challenge is in conjunction with the 13th BEA Workshop and NAACL-HLT 2018 conference.

Introduction

As educational apps increase in popularity, vast amounts of student learning data become available, which can and should be used to drive personalized instruction. While there have been some recent advances in domains like mathematics, modeling **second language acquisition (SLA)** is more nuanced, involving the interaction of lexical knowledge, morpho-syntactic processing, and other skills. Furthermore, most work in NLP for second language (L2) learners has focused on intermediate-to-advanced students of English in assessment settings. Much less work has been done involving beginners, learners of languages other than English, or study over time.

This task aims to forge new territory by utilizing student trace data from users of Duolingo, the world's most popular online language-learning platform. Participating teams are provided with transcripts from millions of exercises completed by thousands of students over their first 30 days of learning on Duolingo. These transcripts are annotated for token (word) level mistakes, and the task is to predict what mistakes each learner will make in the future.

Novel and interesting research opportunities in this task:

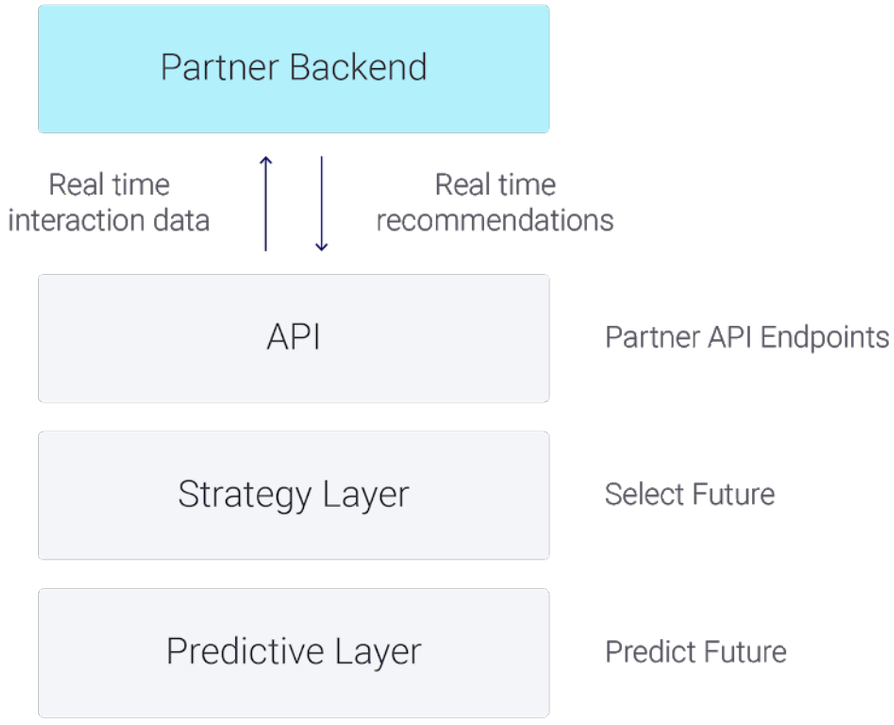
- There will be three (3) tracks for learners of **English**, **Spanish**, and **French**. Teams are encouraged to explore features which generalize across all three languages.
- Anonymized **user IDs** and **time data** will be provided. This allows teams to explore various personalized, adaptive SLA modeling approaches.
- The sequential nature of the data also allows teams to model language **learning** (and **forgetting**) over time.

By accurately modeling student mistake patterns, we hope this task will shed light on both (1) the inherent nature of L2 learning, and (2) effective ML/NLP engineering strategies to build personalized adaptive learning systems.

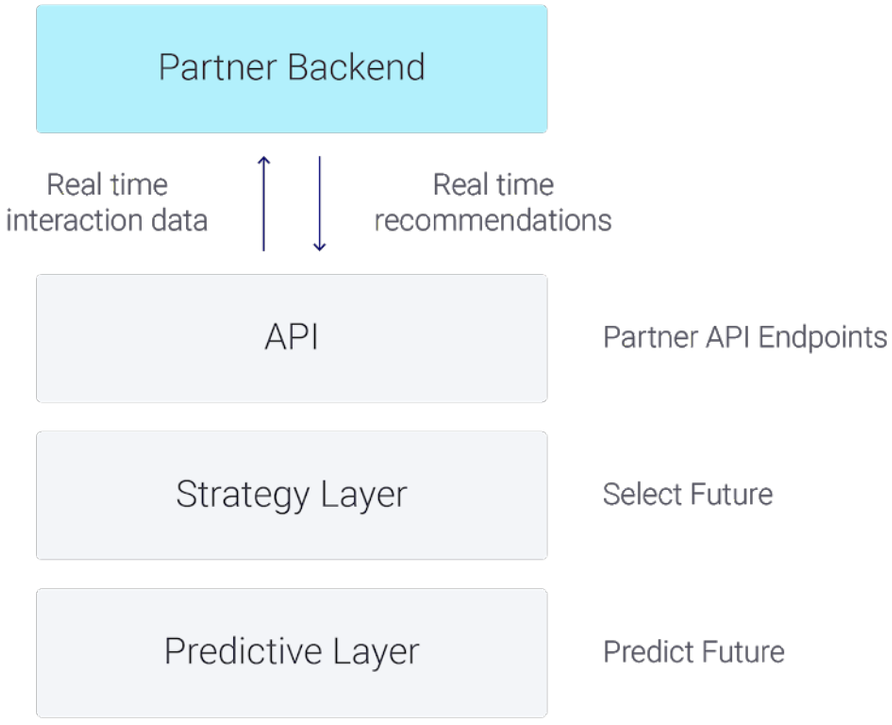
Important Dates

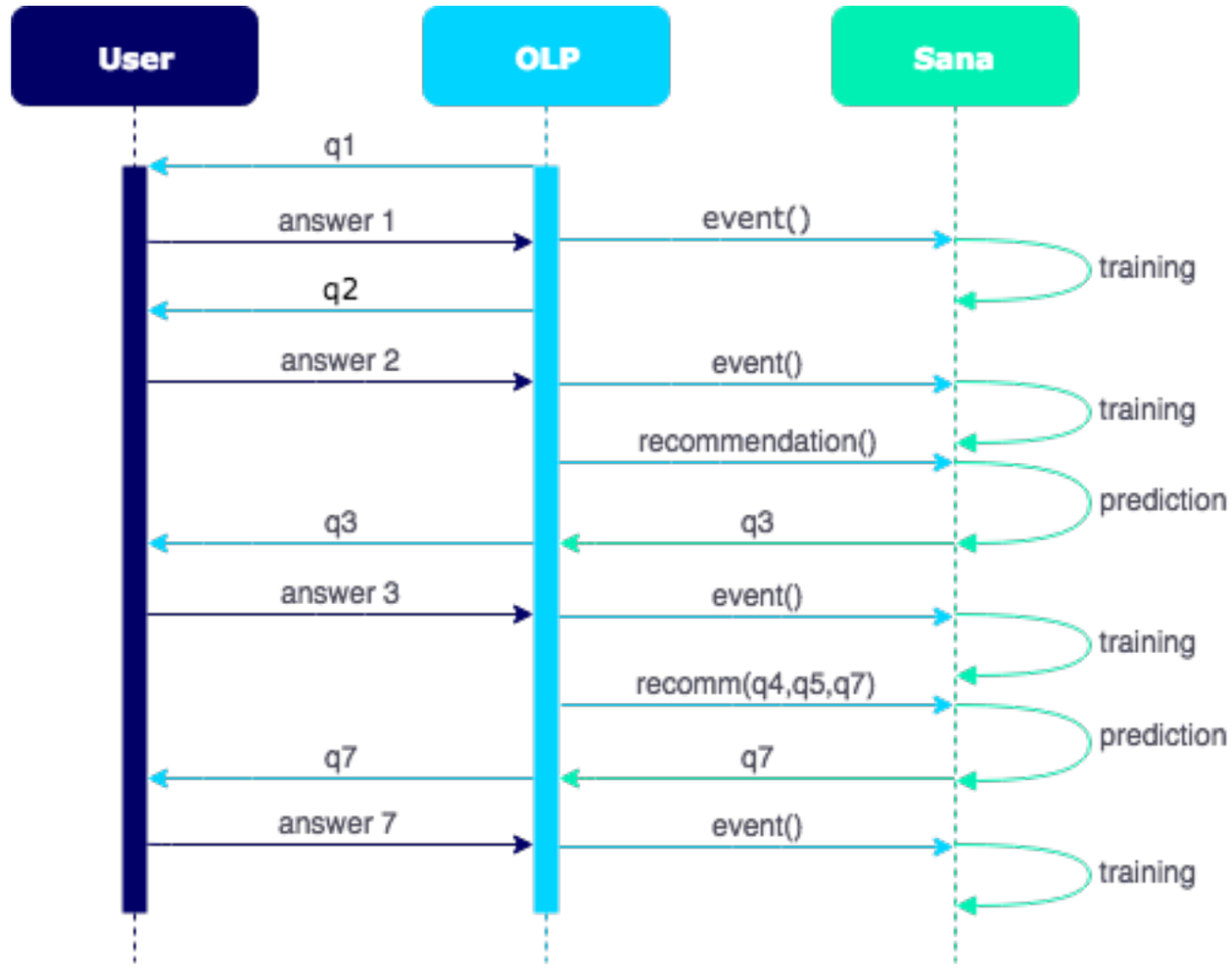
June 05, 2018	Workshop at NAACL-HLT in New Orleans! (Link, Registration)
April 16, 2018	Camera-ready system papers due (Instructions)
April 10, 2018	System paper reviews returned
March 28, 2018	Draft system papers due
March 21, 2018	Final results announcement
March 19, 2018	Final predictions deadline (CodaLab)
March 9, 2018	Data release (phase 2): blind TEST set (Dataverse)
January 10, 2018	Data release (phase 1): TRAIN and DEV sets (Dataverse)

Sana Labs


Partner Backend
Real time interaction data
Real time recommendations
API
Partner API Endpoints
Strategy Layer
Select Future
Predictive Layer
Predict Future

Sana Labs


Partner Backend
Real time interaction data
Real time recommendations
API
Partner API Endpoints
Strategy Layer
Select Future
Predictive Layer
Predict Future


User
OLP
Sana
q1
answer 1
event()
training
q2
answer 2
event()
training
recommendation()
prediction
q3
q3
answer 3
event()
training
recomm(q4,q5,q7)
prediction
q7
q7
answer 7
event()
training

Sana Labs

Main Algorithmic Components

Predict

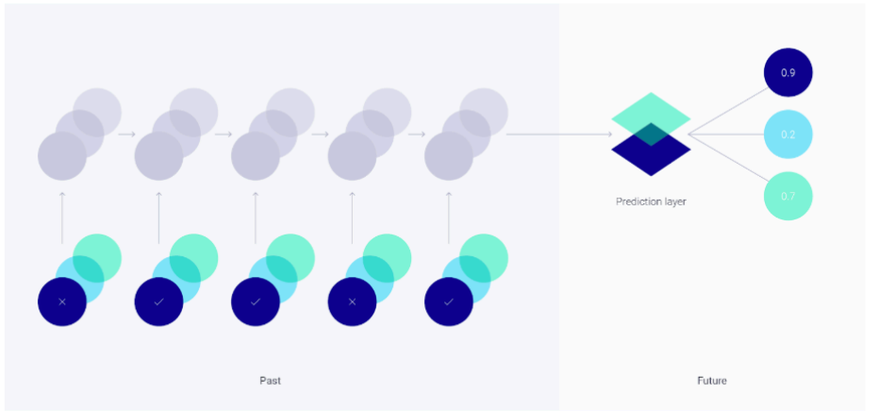
Predict user behaviour

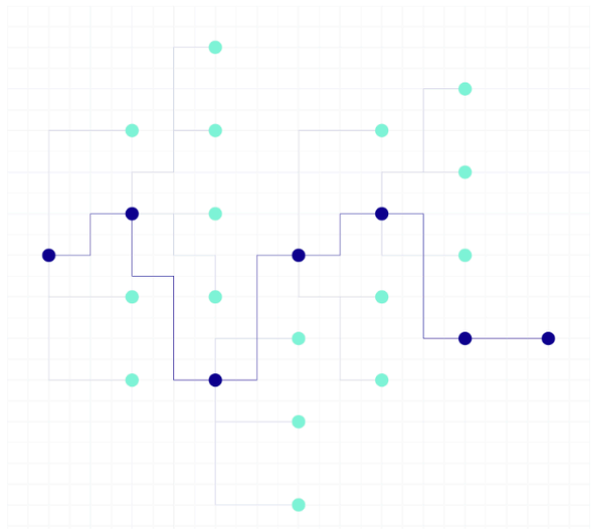
Select

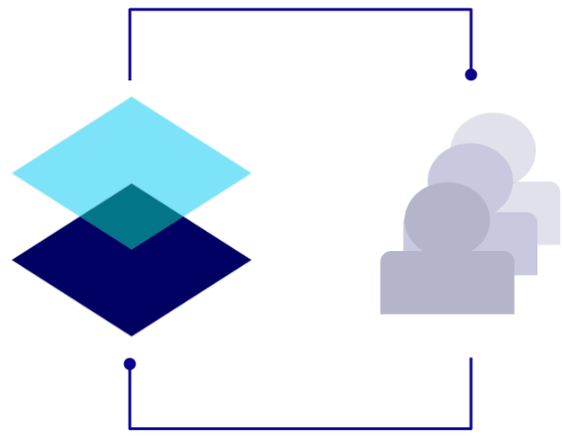
Select optimal content given predictions

Improve

Optimise selection strategies and models towards engagement and learning efficacy





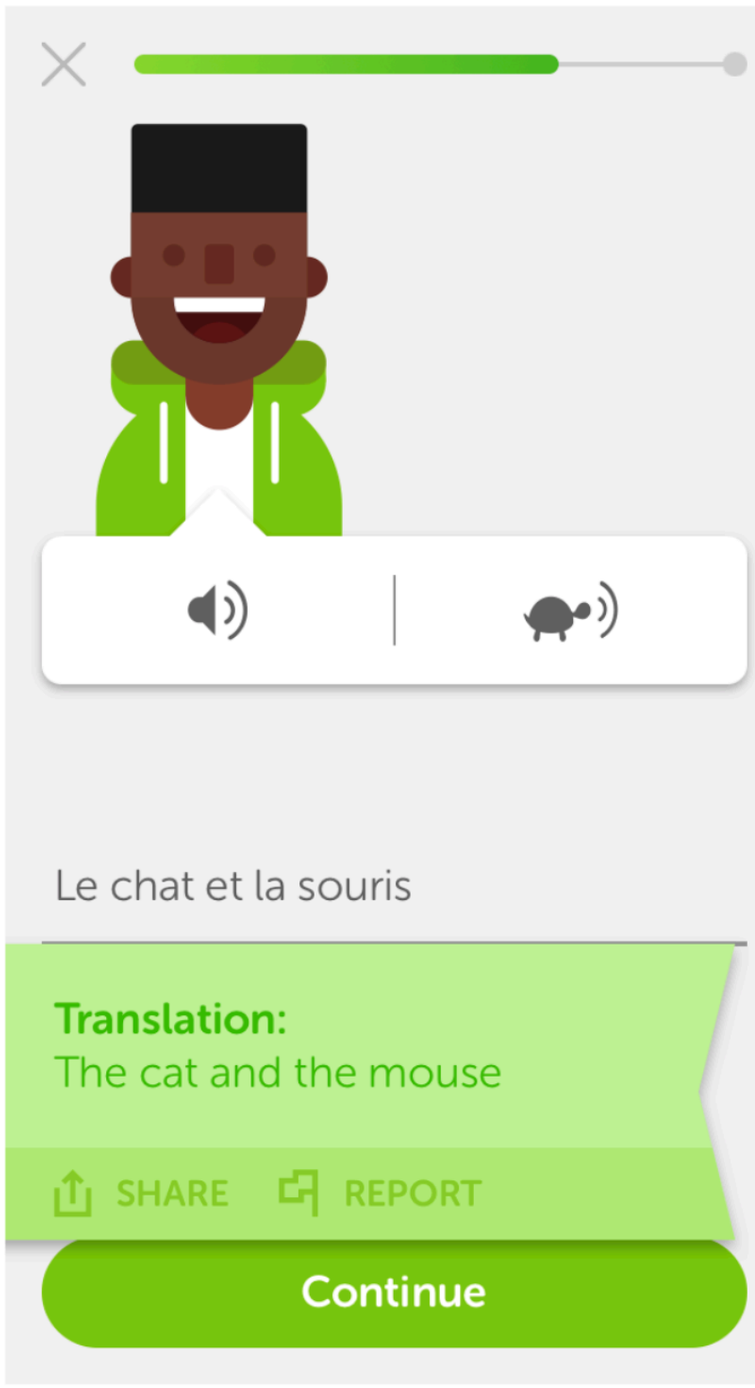


Sana

Personalized education for **everyone**, **everywhere**.

(a) `reverse_translate`

(b) `reverse_tap`

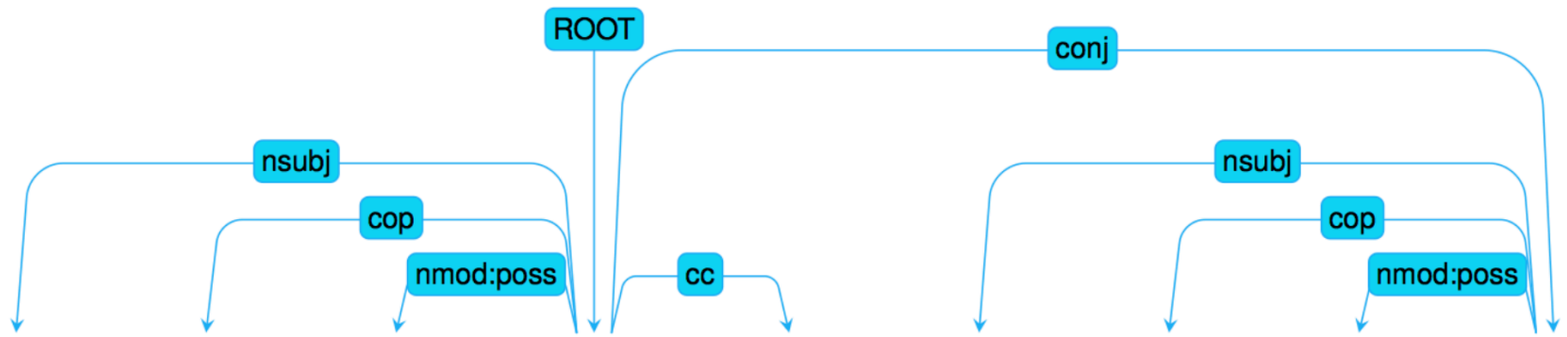


(c) `listen`

TASK: Predict correctness of answered words

Given:

- Metadata (time, device, user’s country, etc)
- Answered words, with linguistic tags and parse tree

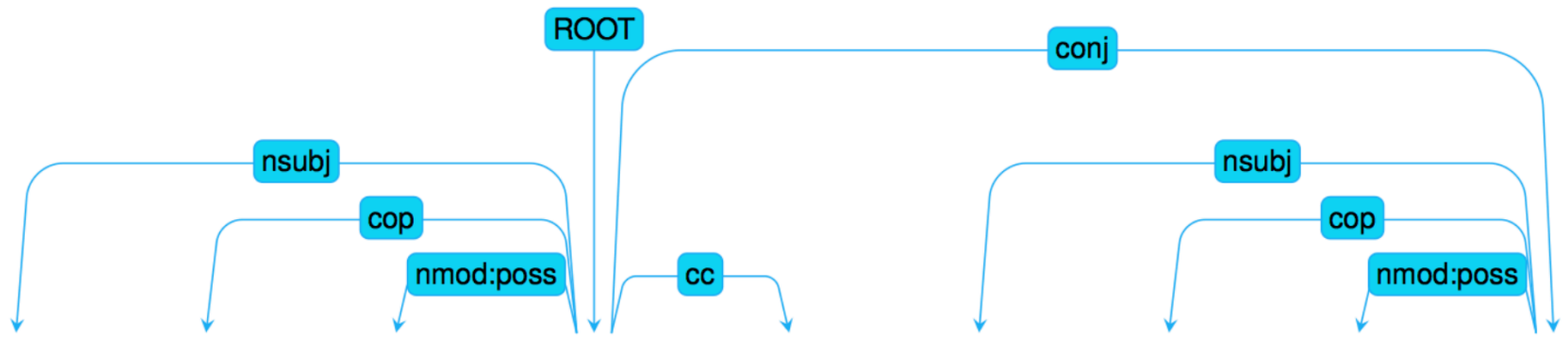


	PRON	VERB	PRON	NOUN	CONJ	PRON	VERB	PRON	NOUN
correct:	She	is	my	mother	and	he	is	my	father
student:	she	is		mader	and	he	is		fhader
label:	O	O	✖	✖	O	O	O	✖	✖

```
# user:D2inSf5+  countries:MX  days:1.793  client:web  session:lesson  format:reverse_translate  time:16
8rgJEAPw1001  She          PRON    Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP    nsubj       4  0
8rgJEAPw1002  is           VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ    cop         4  0
8rgJEAPw1003  my           PRON    Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$              nmod:poss   4  1
8rgJEAPw1004  mother       NOUN    Degree=Pos|fPOS=ADJ++JJ                                                ROOT        0  1
8rgJEAPw1005  and          CONJ    fPOS=CONJ++CC                                                          cc          4  0
8rgJEAPw1006  he           PRON    Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP   nsubj       9  0
8rgJEAPw1007  is           VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ    cop         9  0
8rgJEAPw1008  my           PRON    Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$              nmod:poss   9  1
8rgJEAPw1009  father       NOUN    Number=Sing|fPOS=NOUN++NN                                              conj        4  1

# 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
```

```
# user:D2inSf5+  countries:MX  days:1.793  client:web  session:lesson  format:reverse_translate  time:16
```

```
8rgJEAPw1001  She      PRON  Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP  nsubj      4  0
8rgJEAPw1002  is       VERB  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ  cop        4  0
8rgJEAPw1003  my       PRON  Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$             nmod:poss  4  1
8rgJEAPw1004  mother   NOUN  Degree=Pos|fPOS=ADJ++JJ                                              ROOT       0  1
8rgJEAPw1005  and      CONJ  fPOS=CONJ++CC                                                        cc         4  0
8rgJEAPw1006  he       PRON  Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP  nsubj      9  0
8rgJEAPw1007  is       VERB  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ  cop        9  0
8rgJEAPw1008  my       PRON  Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$             nmod:poss  9  1
8rgJEAPw1009  father   NOUN  Number=Sing|fPOS=NOUN++NN                                             conj       4  1
```



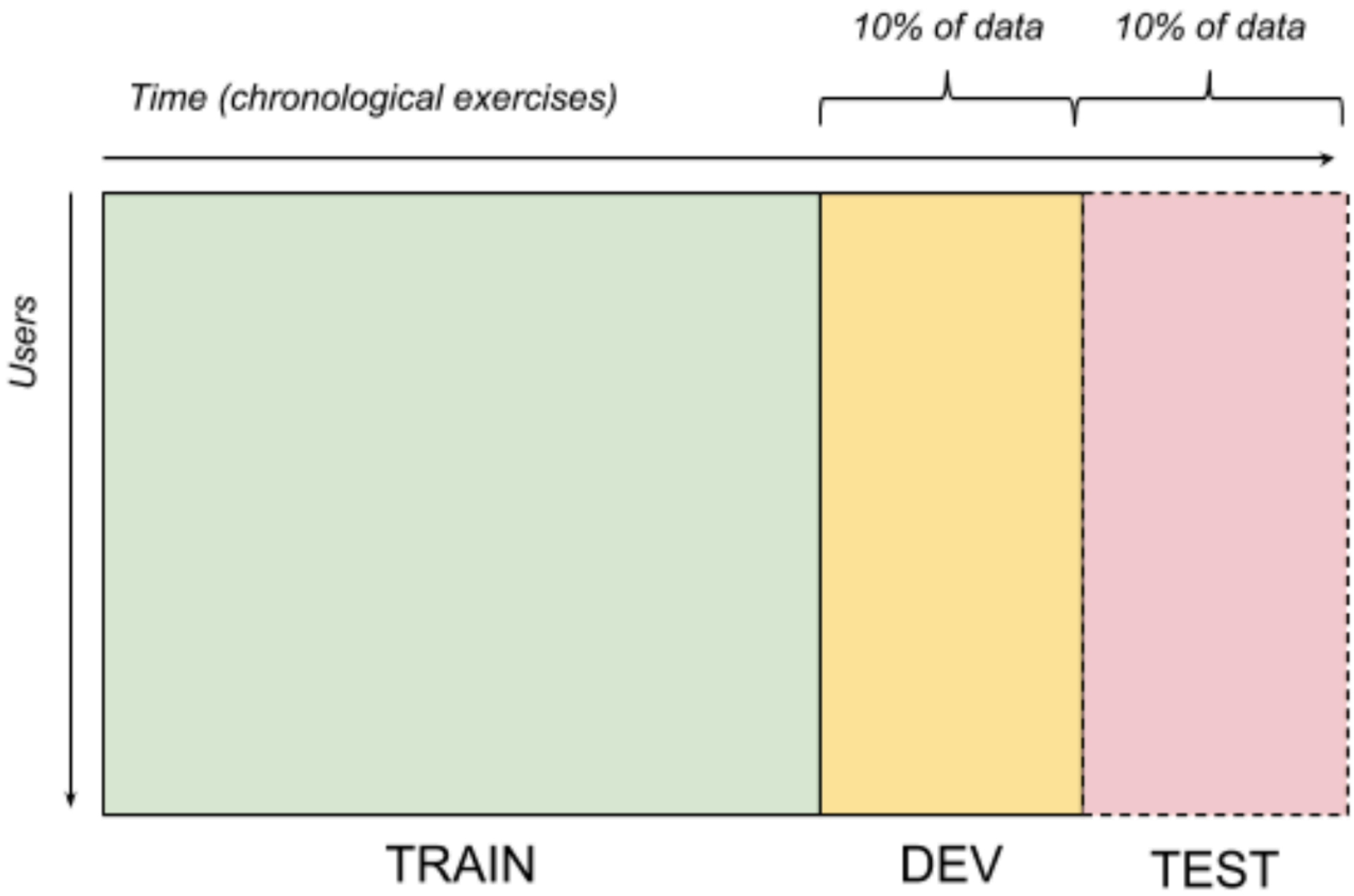
	PRON	VERB	PRON	NOUN	CONJ	PRON	VERB	PRON	NOUN
correct:	She	is	my	mother	and	he	is	my	father
student:	she	is		mader	and	he	is		fhader
label:	O	O	✖	✖	O	O	O	✖	✖

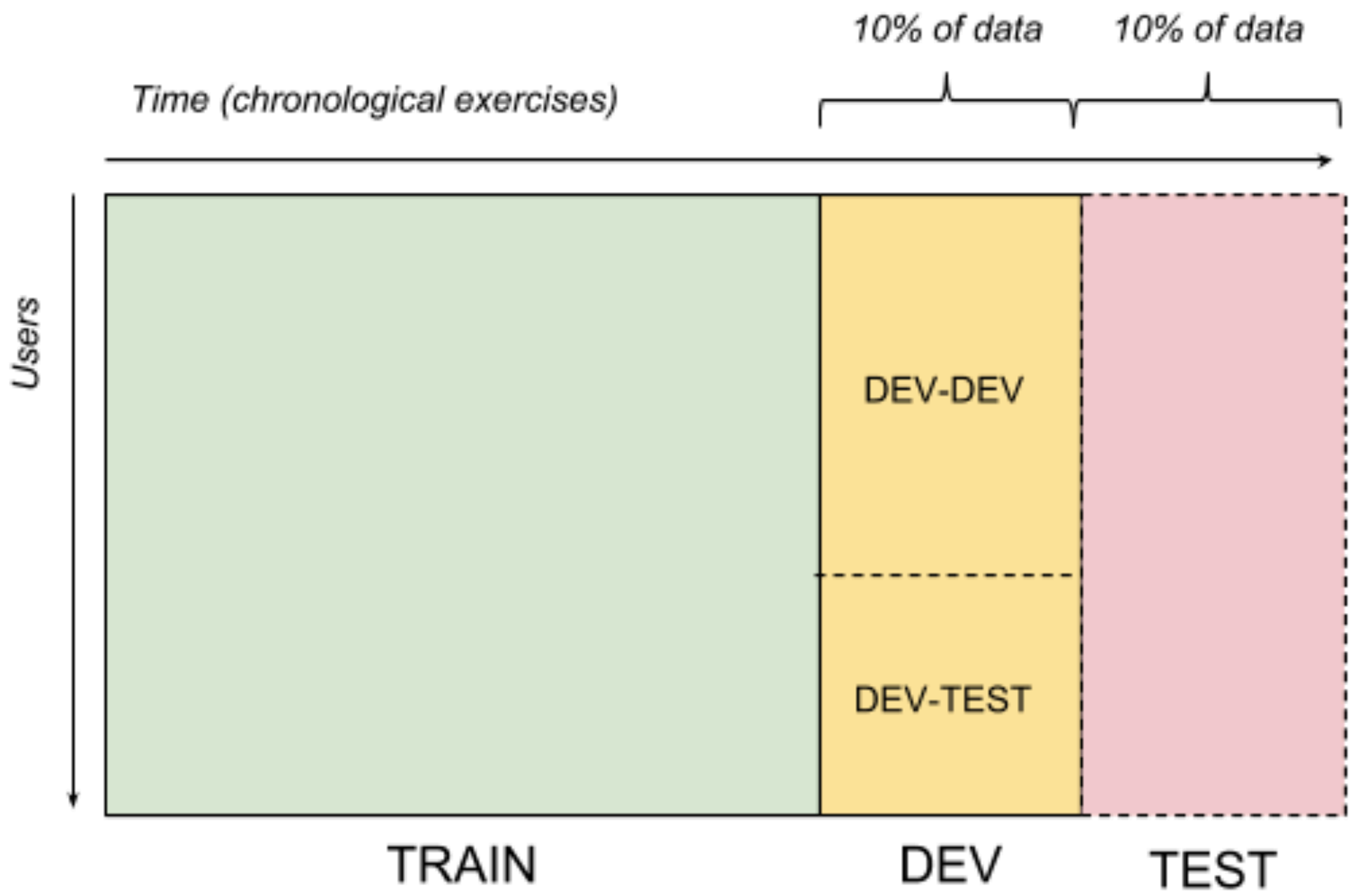
```
# user:D2inSf5+  countries:MX  days:1.793  client:web  session:lesson  format:reverse_translate  time:16
8rgJEAPw1001  She          PRON    Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP    nsubj       4  0
8rgJEAPw1002  is           VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ    cop         4  0
8rgJEAPw1003  my           PRON    Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$              nmod:poss   4  1
8rgJEAPw1004  mother       NOUN    Degree=Pos|fPOS=ADJ++JJ                                                ROOT        0  1
8rgJEAPw1005  and          CONJ    fPOS=CONJ++CC                                                          cc          4  0
8rgJEAPw1006  he           PRON    Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP   nsubj       9  0
8rgJEAPw1007  is           VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ    cop         9  0
8rgJEAPw1008  my           PRON    Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$              nmod:poss   9  1
8rgJEAPw1009  father       NOUN    Number=Sing|fPOS=NOUN++NN                                              conj        4  1

# 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
```

Three datasets

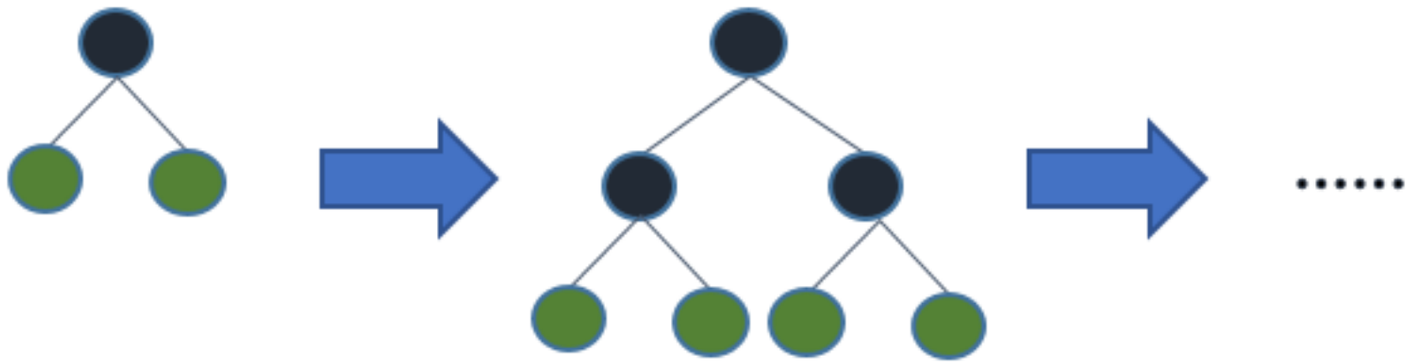
- `en_es` — **English learners**
- `es_en` — **Spanish learners**
- `fr_en` — **French learners**

- For every user, their time ordered interactions were split into train/dev/test
  - First 80% = train
  - Middle 10% = dev
  - Last 10% = test

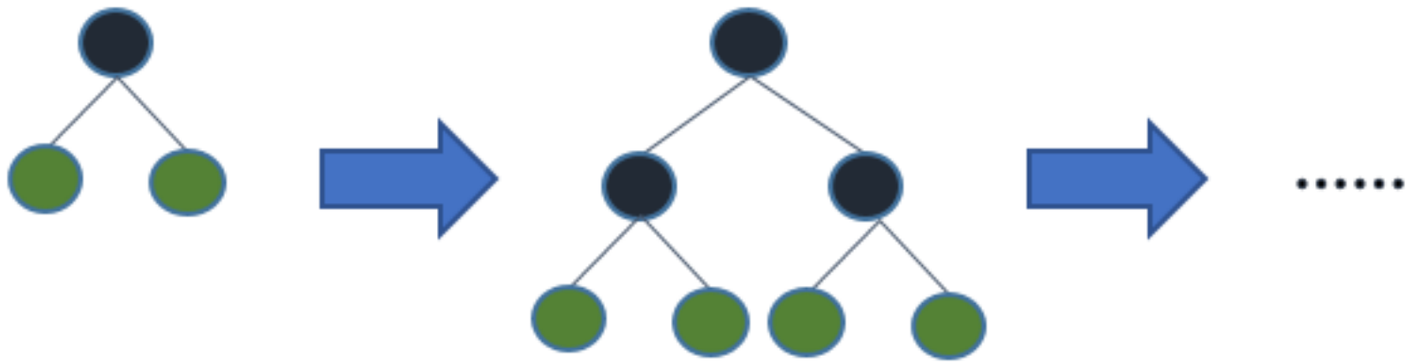

10% of data
10% of data
Time (chronological exercises)
Users
TRAIN
DEV
TEST


10% of data
10% of data
Time (chronological exercises)
Users
DEV-DEV
DEV-TEST
TRAIN
DEV
TEST

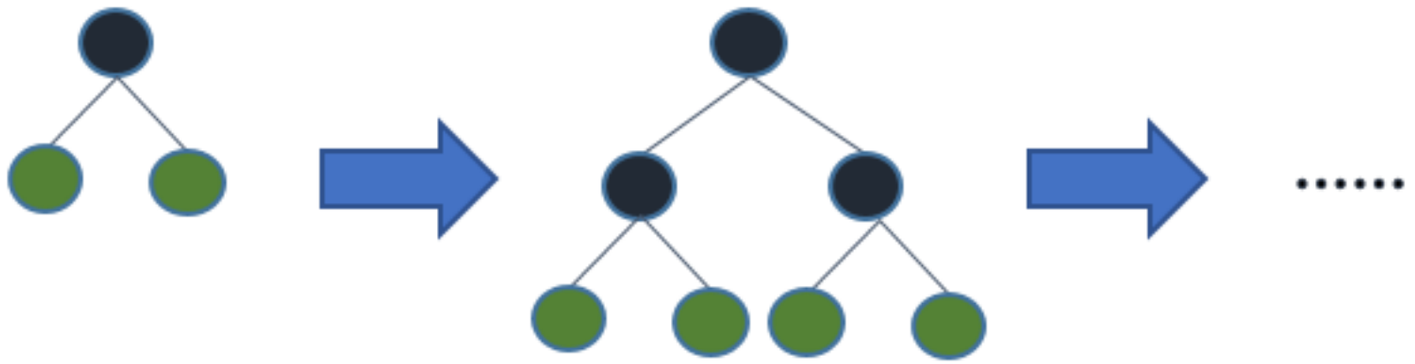
Gradient Boosted Decision Tree Model



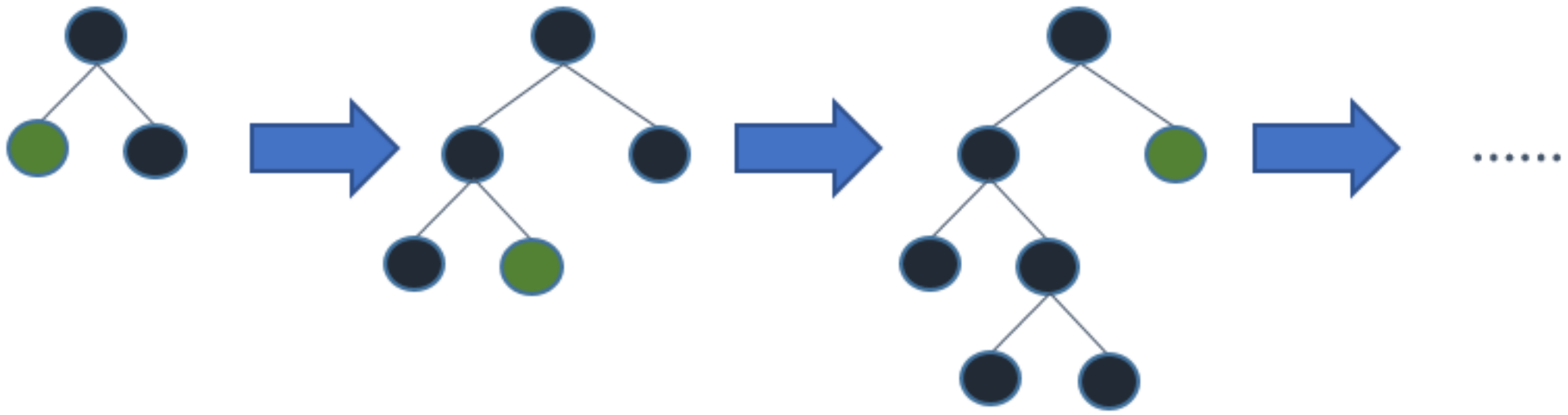
LightGBM framework



Level-wise tree growth



Level-wise tree growth



Leaf-wise tree growth

LightGBM framework

Optimal Split for Categorical Features

Leaf-wise Tree Growth

Binning for Continuous Features

Features

- Combined exercise level and word level features

- Some engineered features

Training

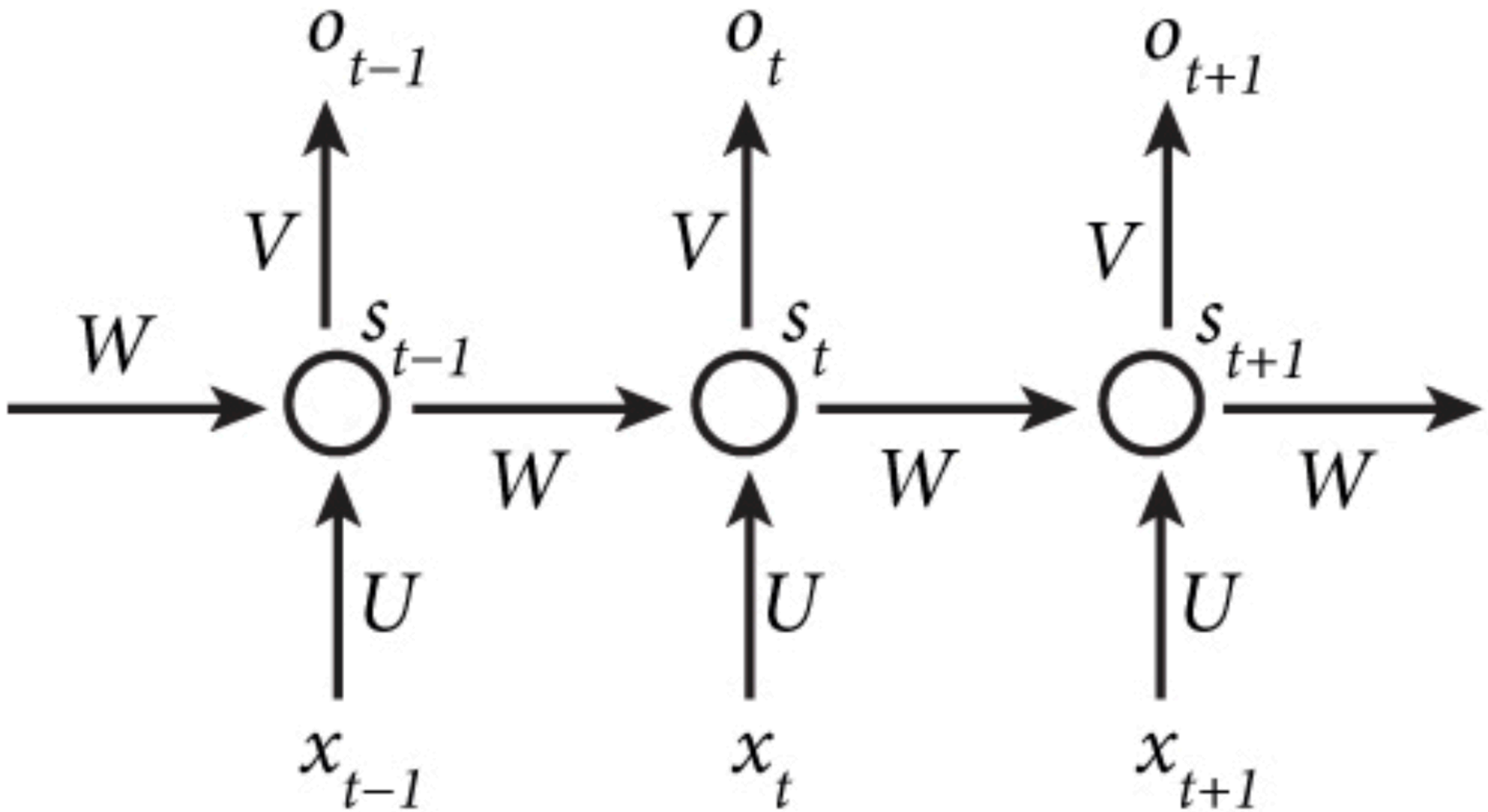
Conclusions:

- Increase number of trees
- Decrease learning rate
- Randomly sample ~40% of features for each tree

Recurrent Neural Network Model

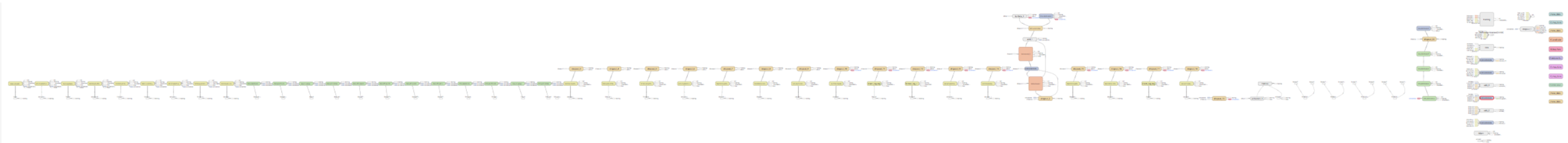
Predictions

Previous State

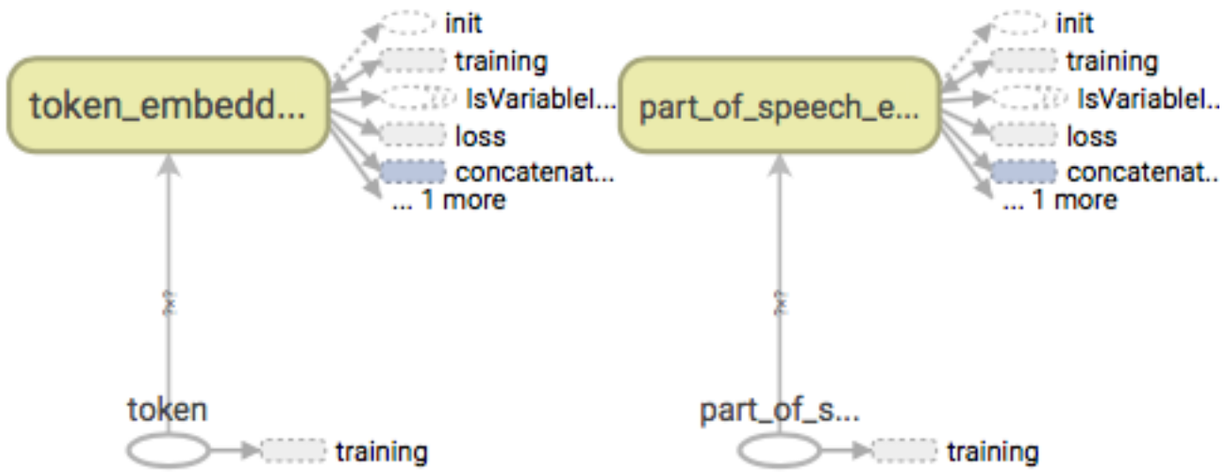


Input

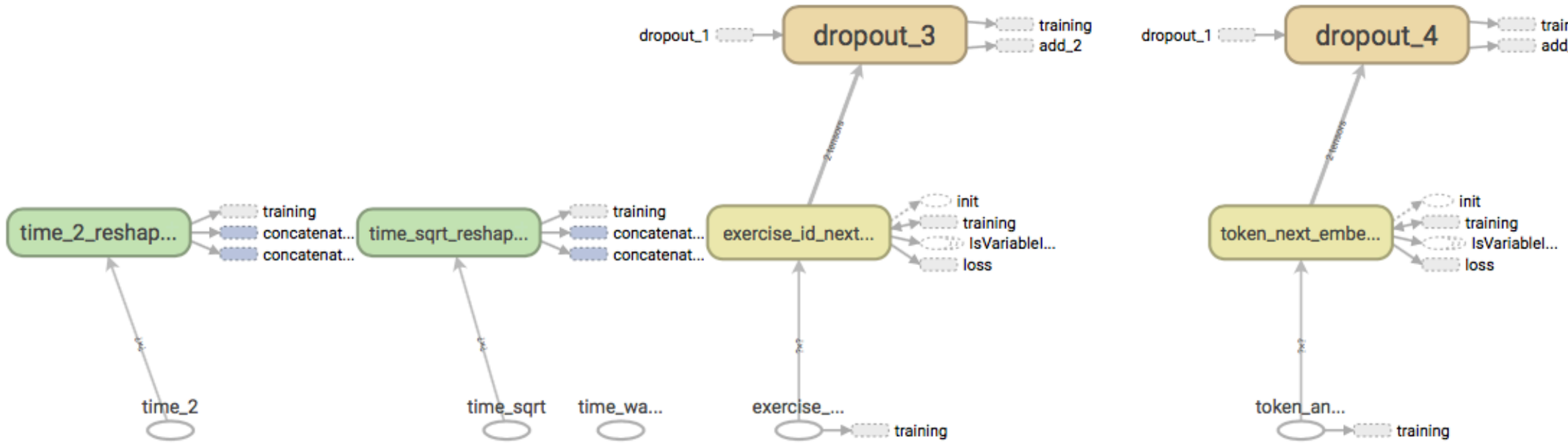
Architecture

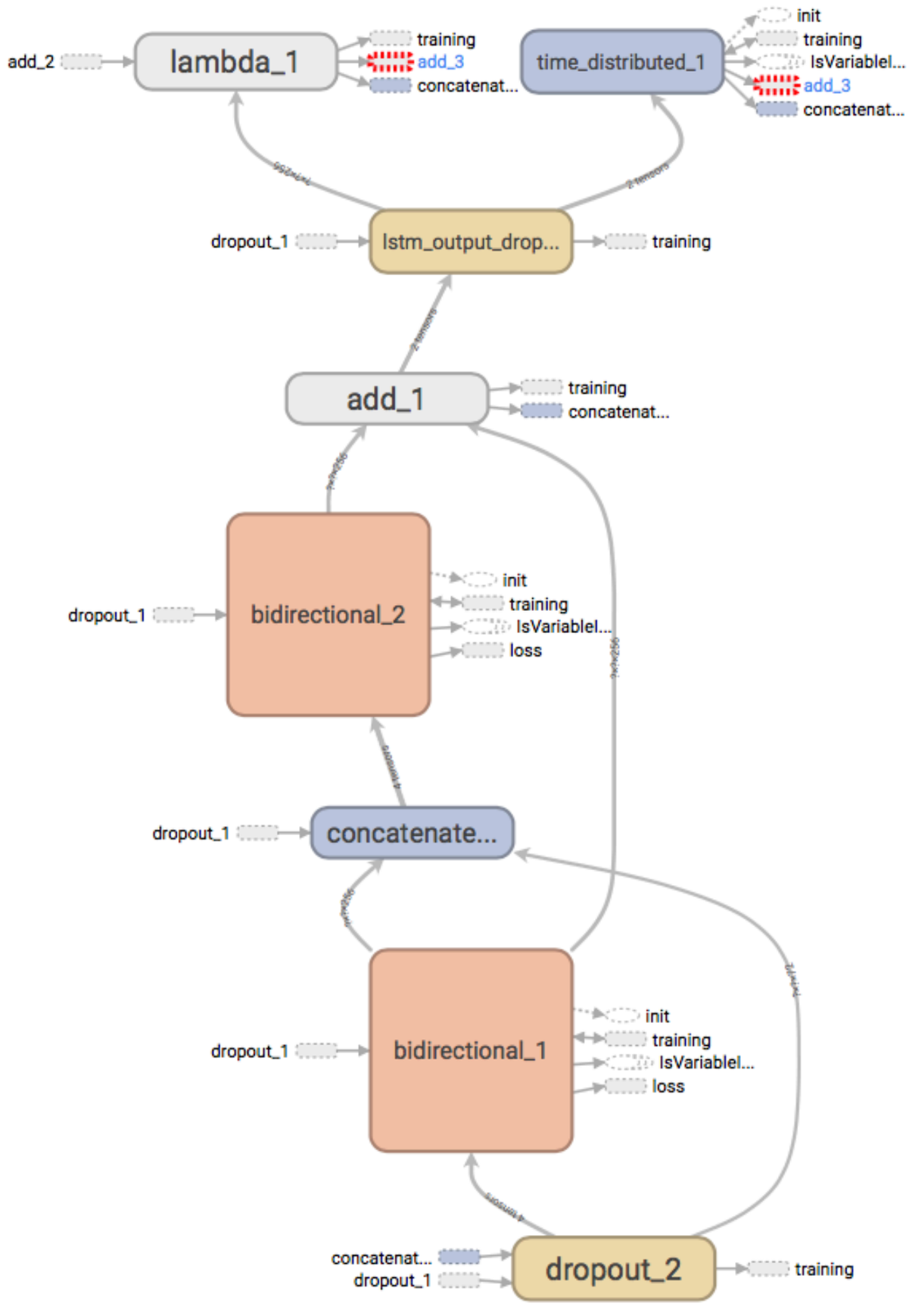


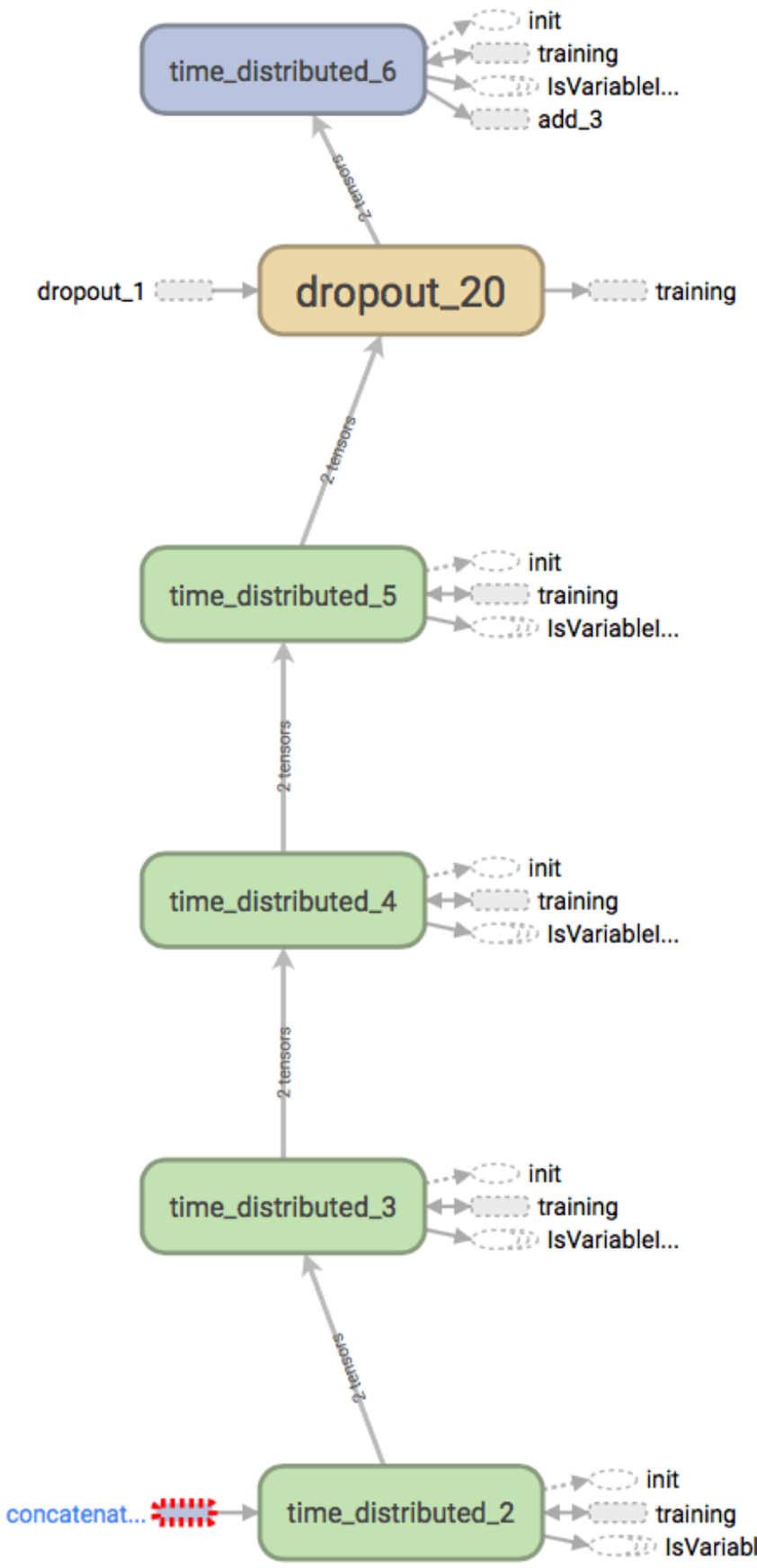
Input


token_embedd...
init
training
IsVariableI...
loss
concatenat...
... 1 more
part_of_speech_e...
init
training
IsVariableI..
loss
concatenat..
... 1 more
int32
int32
token
training
part_of_s...
training

...


dropout_1
dropout_3
training
add_2
dropout_1
dropout_4
training
add_2
time_2_reshap...
training
concatenat...
concatenat...
time_sqrt_reshap...
training
concatenat...
concatenat...
exercise_id_next...
init
training
IsVariableI...
loss
token_next_embe...
init
training
IsVariableI...
loss
int32
int32
int32
time_2
time_sqrt
time_wa...
exercise_...
training
token_an...
training


add_2
lambda_1
training
add_3
concatenat...
time_distributed_1
init
training
IsVariableI...
add_3
concatenat...
dropout_1
lstm_output_drop...
training
add_1
training
concatenat...
dropout_1
bidirectional_2
init
training
IsVariableI...
loss
dropout_1
concatenate...
dropout_1
bidirectional_1
init
training
IsVariableI...
loss
concatenat...
dropout_1
dropout_2
training

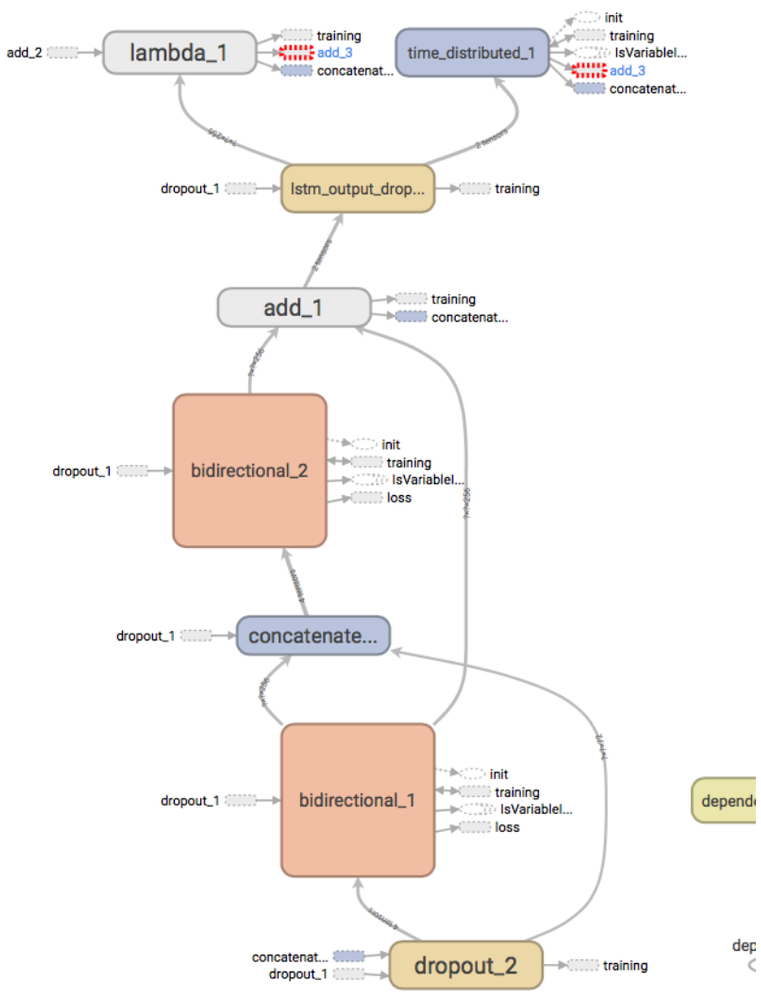

time_distributed_6
init
training
IsVariableI...
add_3
2 tensors
dropout_1
dropout_20
training
2 tensors
time_distributed_5
init
training
IsVariableI...
2 tensors
time_distributed_4
init
training
IsVariableI...
2 tensors
time_distributed_3
init
training
IsVariableI...
2 tensors
concatenat...
time_distributed_2
init
training
IsVariableI...

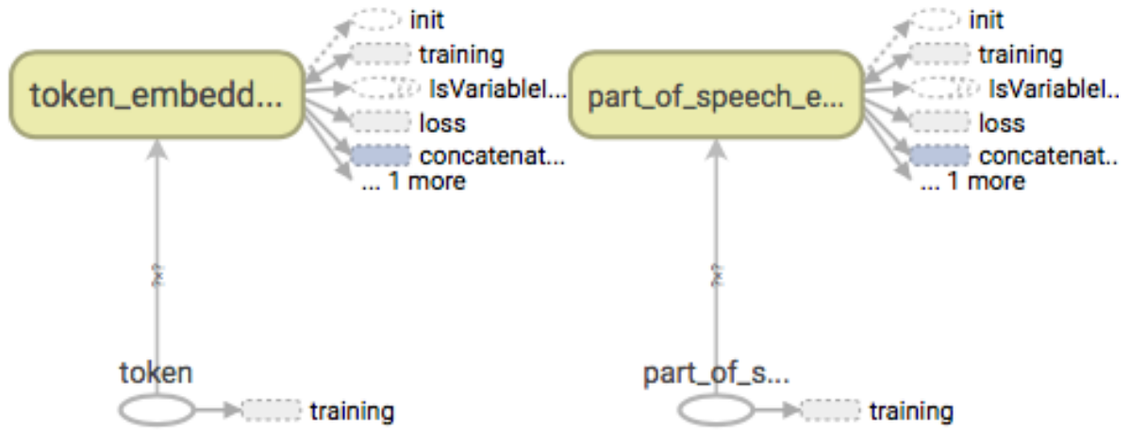
logit term

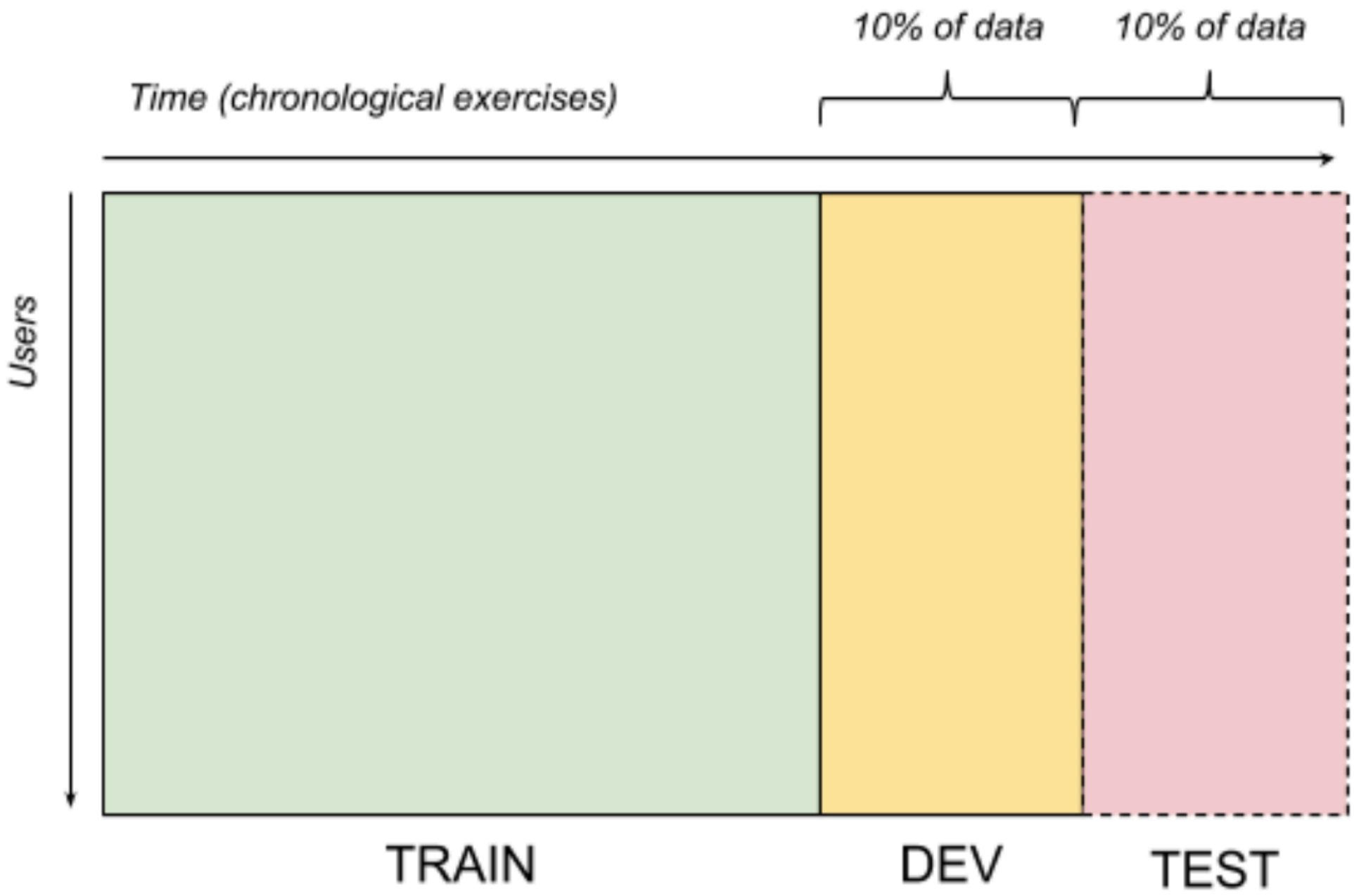
logit term

dot product

concatenate






10% of data
10% of data
Time (chronological exercises)
Users
TRAIN
DEV
TEST

Training

- Sample sequences of 256 words
- Hide labels for second half

Training

- Dropout and L2 Regularisation for embeddings, recurrent and feed forward layers
- Adam with Gradient clipping
- Gaussian Process Bandit Optimisation

Results

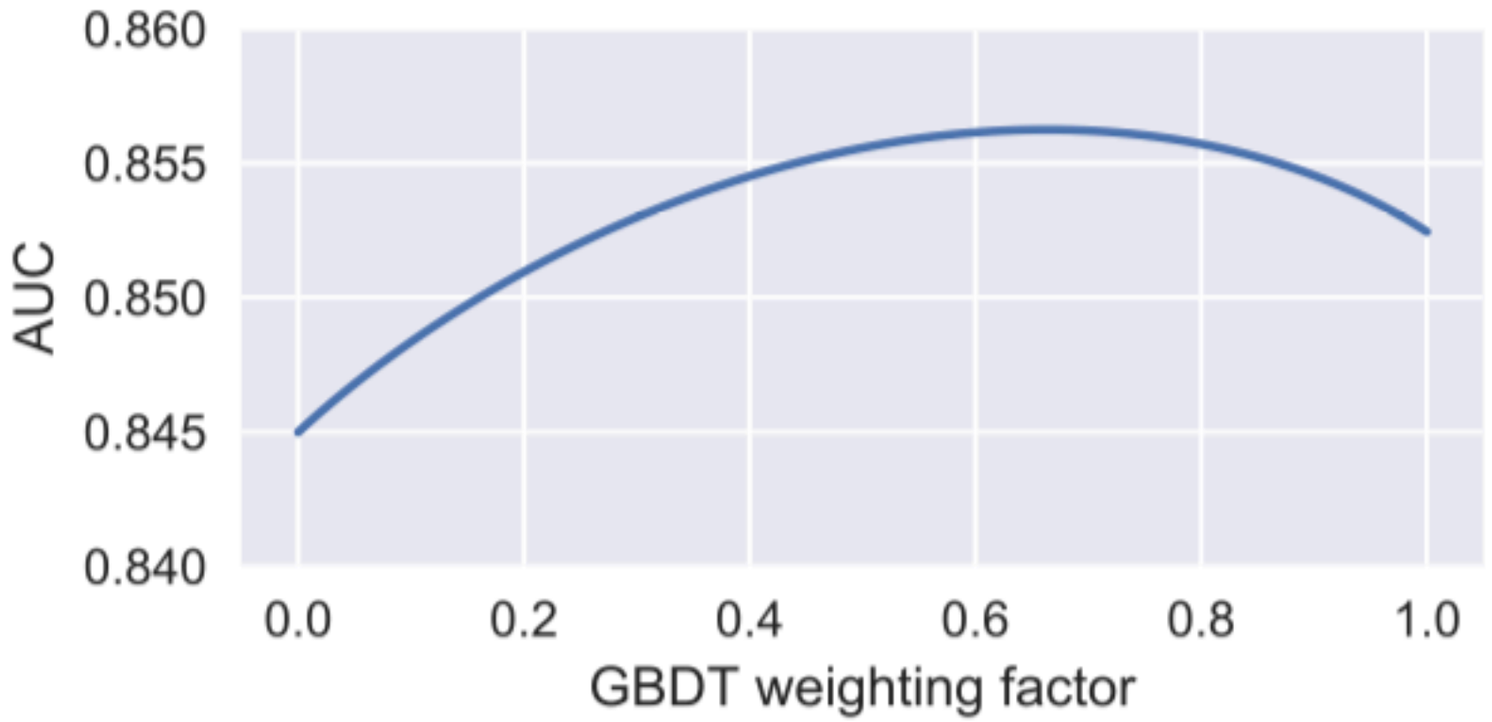
Results

	AUC
GBDT	• 0.8573
RNN	• 0.8506

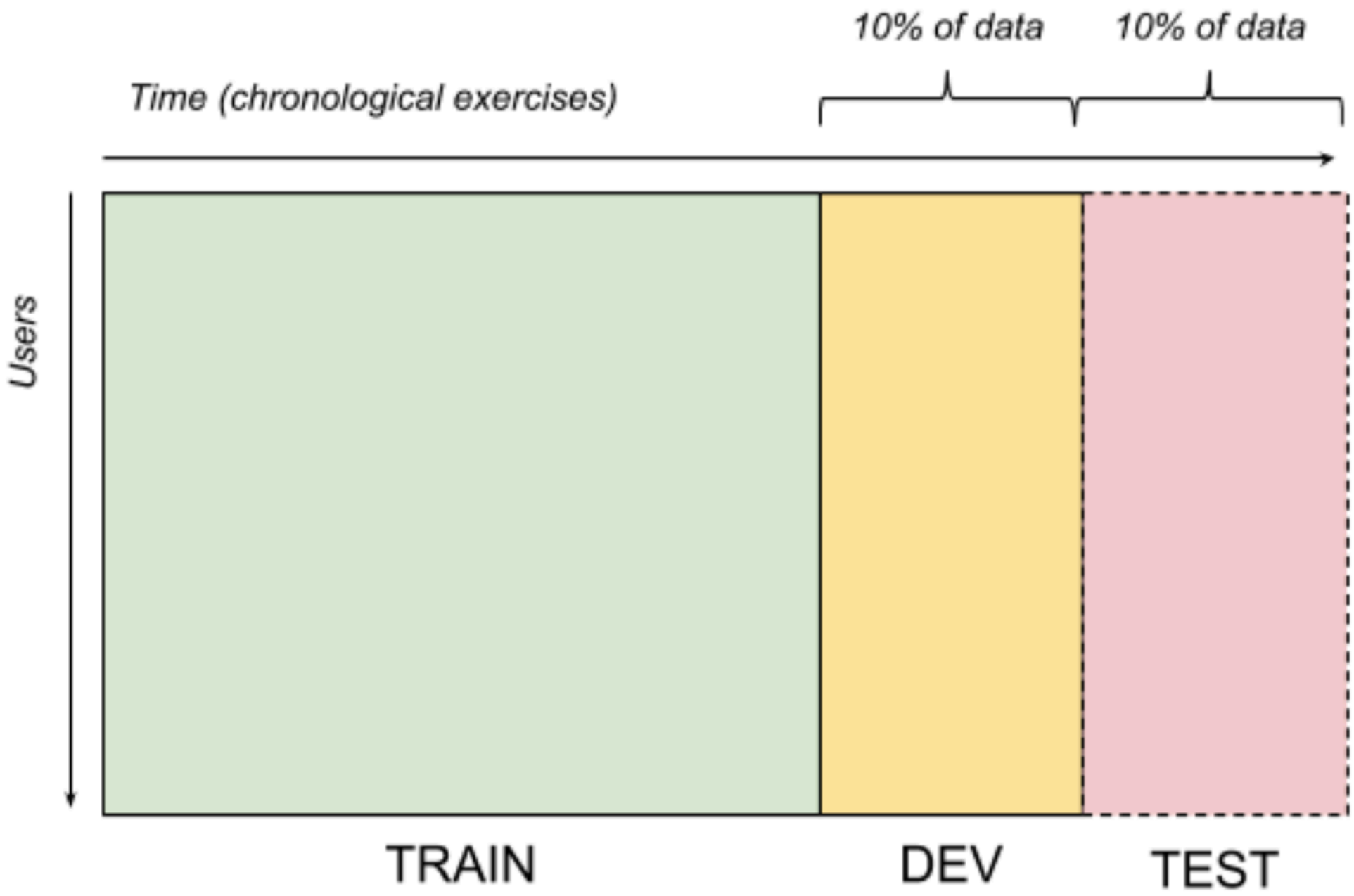
Ensambling

Ensambling

- Pick weighting factor maximising AUC on the dev set
- Retrain models on all data
- Predict on test set


0.860
0.855
0.850
0.845
0.840
AUC
0.0
0.2
0.4
0.6
0.8
1.0
GBDT weighting factor

Analysis


10% of data
10% of data
Time (chronological exercises)
Users
TRAIN
DEV
TEST

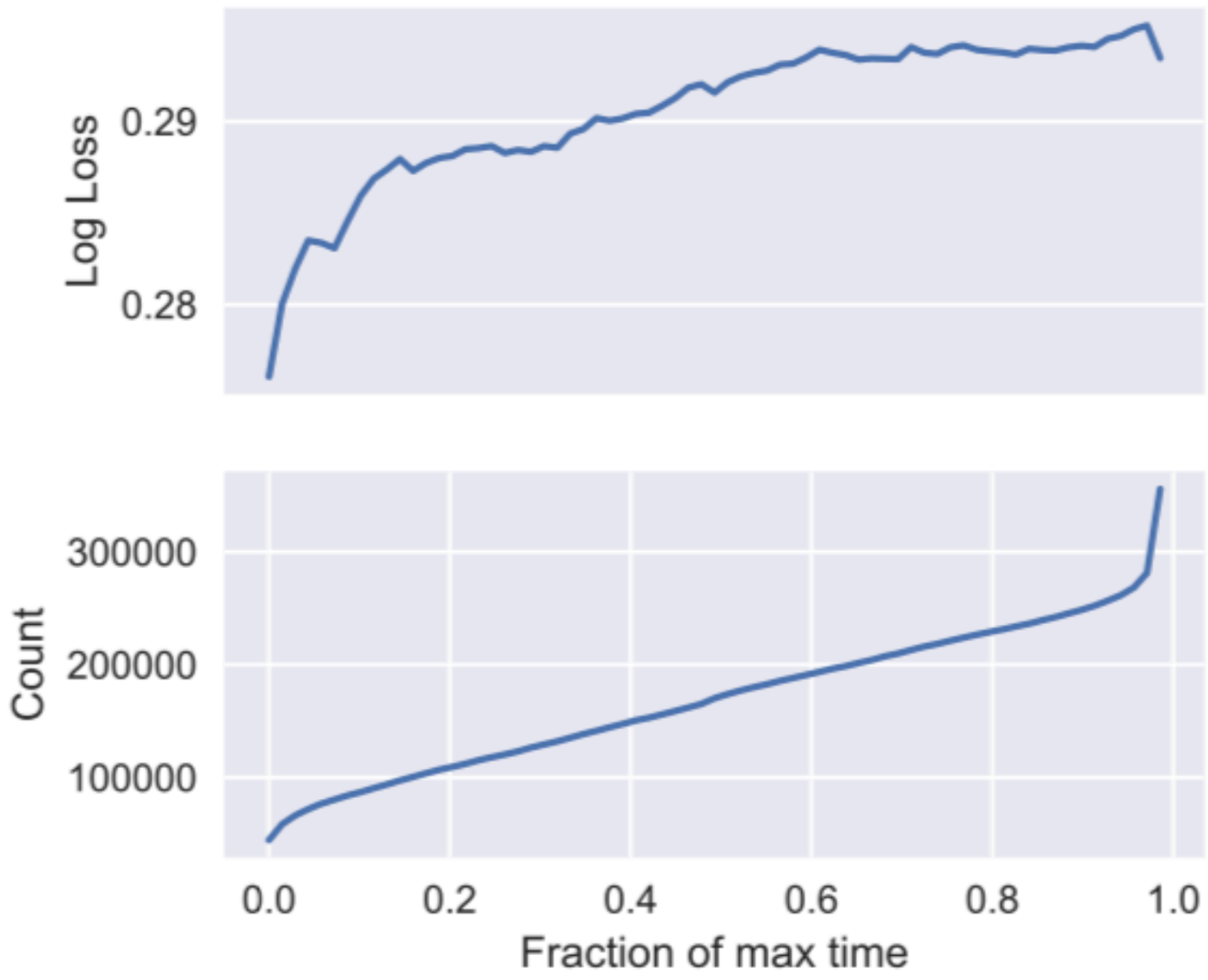


Figure 2: Performance decays as instances further away from the label horizon are considered.

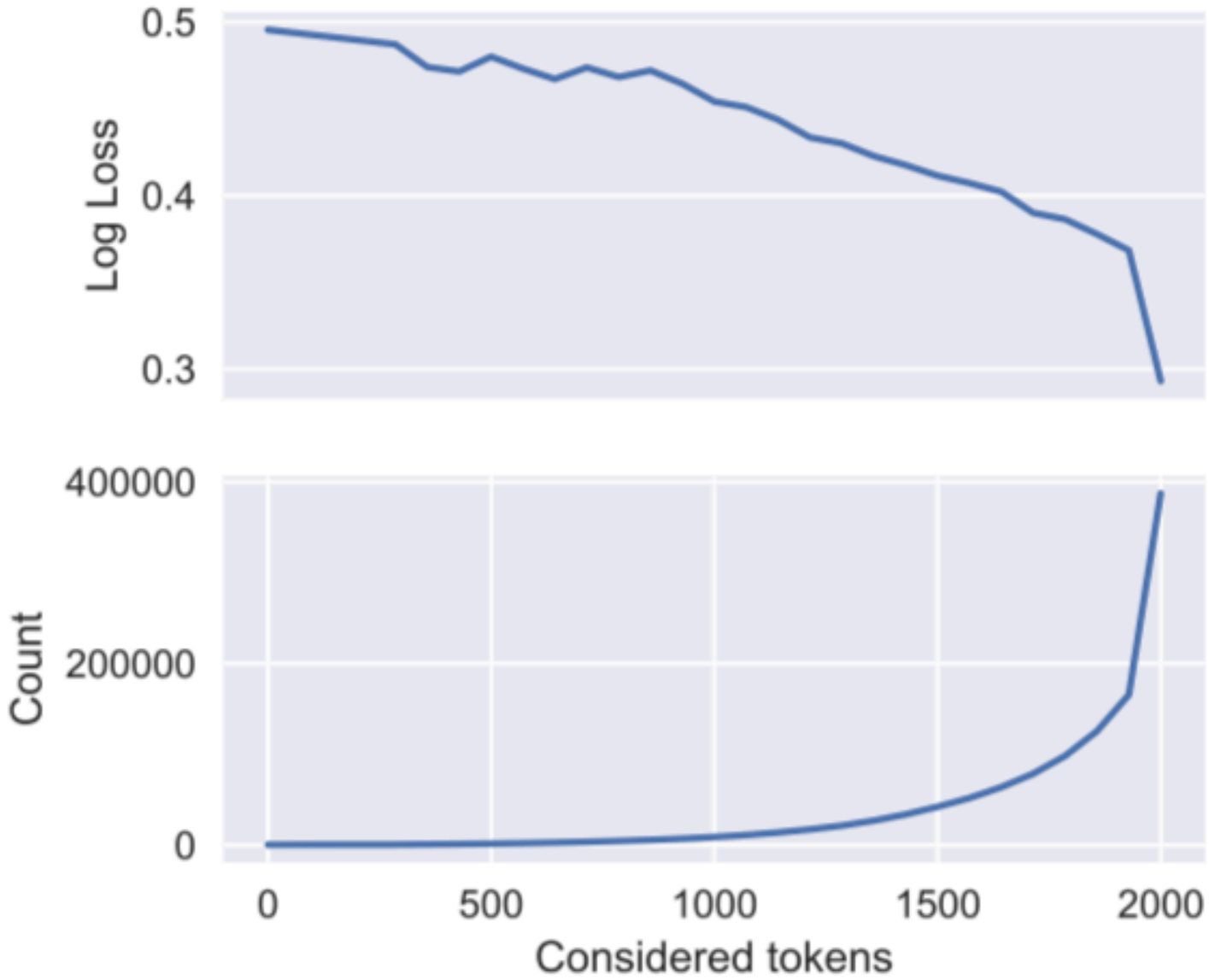


Figure 3: Log loss is high when considering only the x most rare tokens

English Track

↑	Team	AUC	F1
1	SanaLabs ♢♣	.861	.561
1	singsound ♢	.861	.559
3	NYU ♣‡	.859	.468
4	TMU ♢‡	.848	.476
5	CECL ‡	.846	.414
6	Cambridge ♢	.841	.479
7	UCSD ♣	.829	.424
8	nihalnayak	.821	.376
8	LambdaLab ♣	.821	.389
10	Grotoco	.817	.462
11	jilljenn	.815	.329
12	ymatusevych	.813	.381
13	renhk	.797	.448
14	zlb241	.787	.003
15	SLAM_baseline	.774	.190

Spanish Track

↑	Team	AUC	F1
1	SanaLabs ♢♣	.838	.530
2	NYU ♣‡	.835	.420
2	singsound ♢	.835	.524
4	TMU ♢‡	.824	.439
5	CECL ‡	.818	.390
6	Cambridge ♢	.807	.435
7	UCSD ♣	.803	.375
7	LambdaLab ♣	.801	.344
9	Grotoco	.791	.452
9	nihalnayak	.790	.338
11	ymatusevych	.789	.347
11	jilljenn	.788	.306
13	renhk	.773	.432
14	SLAM_baseline	.746	.175
15	zlb241	.682	.389

French Track

↑	Team	AUC	F1
1	SanaLabs ♢♣	.857	.573
2	singsound ♢	.854	.569
2	NYU ♣‡	.854	.493
4	CECL ‡	.843	.487
5	TMU ♢‡	.839	.502
6	Cambridge ♢	.835	.508
7	UCSD ♣	.823	.442
8	LambdaLab ♣	.815	.415
8	Grotoco	.813	.502
10	nihalnayak	.811	.431
10	jilljenn	.809	.406
10	ymatusevych	.808	.441
13	simplelinear	.807	.394
14	renhk	.796	.481
15	SLAM_baseline	.771	.281

Table 2: Final results. Ranks (↑) are determined by statistical ties (see text). Markers indicate which systems include recurrent neural architectures (♢), decision tree ensembles (♣), or a multitask model across all tracks (‡).

Second Language Acquisition Modeling: An Ensemble Approach

Anton Osika, Susanna Nilsson, Andrii Sydorchuk, Faruk Sahin, Anders Huss
Sana Labs, Nybrogatan 8, 114 34 Stockholm, Sweden
{anton, susanna, andrii, faruk, anders}@sanalabs.com

Abstract

Accurate prediction of students knowledge is a fundamental building block of personalized learning systems. Here, we propose a novel ensemble model to predict student knowledge gaps. Applying our approach to student trace data from the online educational platform Duolingo we achieved highest score on both evaluation metrics for all three datasets in the 2018 Shared Task on Second Language Acquisition Modeling. We describe our model and discuss relevance of the task compared to how it would be setup in a production environment for personalized education.

1 Introduction

Understanding how students learn over time holds the key to unlock the full potential of adaptive learning. Indeed, personalizing the learning experience, so that educational content is recommended based on individual need in real time, promises to continuously stimulate motivation and the learning process (Bauman and Tuzhilin, 2014a). Accurate detection of students' knowl-

2 Data and Evaluation Setup

The 2018 Shared Task on SLAM provides student trace data from users on the online educational platform Duolingo (Settles et al., 2018). Three different datasets are given representing users responses to exercises completed over the first 30 days of learning English, French and Spanish as a second language. Common for all exercises is that the user responds with a sentence in the language learnt. Importantly, the raw input sentence from the user is not available but instead the best matching sentence among a set of correct answer sentences. The prediction task is to predict the word-level mistakes made by the user, given the best matching sentence and a number of additional features provided. The matching between user response and correct sentence was derived by the finite-state transducer method (Mohri, 1997).

All datasets were pre-partitioned into training, development and test subsets, where approximately the last 10 % of the events for each user is used for testing and the last 10 % of the remaining events used for development . Target labels for

anton.osika@gmail.com