Case Study

State of the Art Modeling of Language Learning for Duolingo

The complexity of personalizing language learning lies in its nuances, involving the interaction of lexical knowledge, morpho-syntactic processing, human memory, and other aspects. The emergence of large datasets combined with rich models is now empowering us to capture these nuances to model and personalize language learning more effectively.

Problem

Since its founding, Duolingo has empowered 300 million students across the globe. As students learn with Duolingo, they create immense amounts of data on their various mistakes along the way. In analysing these mistakes, student knowledge gaps can be detected. The goal of this work was to predict future mistakes that learners of English, Spanish, and French will make based on the mistakes they have made in the past.

Solution

The problem is inherently difficult as language learning is grounded in the complexity of both the human brain and human knowledge. Hence, the use of rich models is appropriate. Our system combines a Recurrent Neural Network (RNN) and a Gradient Boosted Decision Tree (GBDT). This approach has important advantages over previously employed models of language learning in that it does not require encoding domain knowledge and can capture more nuances of learning.

For example, the student below seems to be struggling with "my", "mother", and "father." These mistakes could point to difficulties with possessive pronouns or the orthography of English ”th” sounds. Our system picks up on these trends, predicting student knowledge gaps in a personalized way that evolves over time. These predictions can subsequently be used to personalize learning by recommending exactly what to study and when, based on each individual’s needs.

	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	✓	✓	✗	✗	✓	✓	✓	✗	✗

Results

Our system performed extraordinarily well, producing the best evaluation metrics on Duolingo’s benchmark. Such results illustrate how machine learning can transcend the limits of traditional approaches. We are excited for this work to impact learners and provide a foundation for further research on knowledge acquisition within all subjects, beyond language learning[1].

TEAM	RANK
SanaLabs	1.0
singsound	1.7
NYU	2.3
TMU	4.3
CECL	4.7
Cambridge	6.0

[1] A. Osika, S. Nilsson, A. Sydorchuk, F. Sahin, A. Huss, "*Second Language Acquisition Modeling: An Ensemble Approach*", Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 217–222, June 2018.