Context Based Approach for Second Language Acquisition

Nihal V. Nayak
Stride.AI
Bangalore, India
nihalnayak@gmail.com

Arjun R. Rao
Ramaiah Institute of Technology
Bangalore, India
mailarjunrao@gmail.com

Abstract

SLAM 2018 focuses on predicting a student's mistake while using the Duolingo application. In this paper, we describe the system we developed for this shared task. Our system uses a logistic regression model to predict the likelihood of a student making a mistake while answering an exercise on Duolingo in all three language tracks - English/Spanish (en/es), Spanish/English (es/en) and French/English (fr/en). We conduct an ablation study with several features during the development of this system and discover that context based features play a major role in language acquisition modeling. Our model beats Duolingo's baseline scores in all three language tracks (AUROC scores for en/es = 0.821, es/en = 0.790 and fr/en = 0.812). Our work makes a case for providing favourable textual context for students while learning second language.

1 Introduction

The SLAM 2018 Shared Task is primarily centered around modeling second language acquisition (Settles et al., 2018) of non-native learners of English, Spanish and French. In this shared task, the principal tool used to assess learners is via Duolingo, one of the world's most popular online learning platforms. The data provided as part of the shared task is collected from the way thousands of students performed in over 4 million exercises during their first 30 days on Duolingo. This data consists of annotations at a word level - that indicate errors made by the user in a particular exercise. The task here is to predict mistakes that a learner is likely to make in future, by building a model from the training dataset given. Such a system would thus be able to model the second language acquisition capabilities of non-native learners of these languages.

In this paper, we present our attempt at modelling second language acquisition, primarily by considering context based features. Using these features our system implements a logistic regression model based on the additive conjugate model (Cen et al., 2008) that considers both the instance level features and user's latent ability, that results in reasonably good performance across the three languages being considered.

The rest of this paper is organized as follows. Section 2 highlights some of the existing research in modelling second language acquisition, that we have considered while developing the system. In Section 3, we discuss the features used in our model. We then present our model along with a few alternative approaches we considered (Section 4). An evaluation of our model on the Development and Test datasets is described in Section 5. Finally, scope for future work is discussed in section 6 and we present our conclusions in Section 7.

2 Related work

The process of learning has been thoroughly studied over the years. The forgetting curve (Ebbinghaus, 1885) has been central to these studies, which posits that memory decays exponentially with time. Research suggests that learning a concept in spaced interval helps in long term retention.

Leitner (1972) proposed a strategy (called Leitner's system) which incorporates spaced learning in flashcards. The system accounts for the student's performance and schedules the learning sessions with the help of buckets. For instance, if the student correctly answers the flashcard, it gets promoted to a higher bucket, thereby more spacing is provided between the learning sessions and if the student incorrectly answers the flashcard, then it

gets demoted to a lower bucket and thus reduces the spacing. Duolingo also implements a variant of the Leitner's system by organizing the cards in virtual buckets.

Apart from the Spacing Effect, there are other theories that have been around for sometime. Experiments by Roediger and Karpicke (2006) indicate that repeated testing increases long term retention. Nayak et al. (2017) developed a flashcard based application which implements the testing effect. They also collect a range of attributes or datapoints (both implicit and explicit data points) from the users.

Data collected from the users can be used for language acquisition modeling. For instance, Duolingo implements HLR (Settles and Meeder, 2016) to implement a trainable model for the forgetting curve. With their model, they attempt to predict the probability of a user correctly recalling a word. In this shared task, the organizers have released a similar dataset.

We posit that, the zone of proximal development (Vygotsky and Cole, 1978) plays a crucial role in language acquisition. The theory suggests that, when a student is in her zone of proximal development, providing appropriate assistance will enable her to complete the task. In language learning, the task is to answer the target word or the exercise given the surrounding words or the context. Therefore, in our work we focus on context based features and explore its effect while answering an exercise.

We use insights from recent works in L2 acquisition from code-switched text as they have focused on learning from context. Labutov and Lipson (2014) carry out experiments to determine the guessability of a word in code switched text. A similar work by Knowles et al. (2016) discuss the factors that can potentially affect the guessability of a German word with English context. We extend these works to model acquisition in multiple languages: English-Spanish, Spanish-English and French-English. For modeling the language acquisition, we make use of an additive conjugate model (Cen et al., 2008), in which we account for both instance level features such as token, part of speech, etc as well as the user's ability. We describe our model in detail in the next sections.

3 Features

In this section, we describe the features we consider in our experiments.

We start looking at the different attributes present in the dataset. These features are selected based on our intuition and past work. For simplicity, we divide the features in 2 categories - baseline features and context features.

3.1 Baseline Features

- **Token (T)** - We preprocess this feature by converting the token to lowercase and store the token as a categorical feature.
- **Part-of-Speech (POS)** - The dataset provides POS information for each token in Universal Dependency format. We use the same POS information without any preprocessing in our model.
- **Morphological Features (M)** - The dataset provides a detailed list of morphological features in Universal Dependency format. We encode each of these features in a separate hash bucket and use it in our model.
- **Dependency Label (D)** - The dataset provides dependency label for each token computed using the language agnostic dependency parser in Google's Syntaxnet.
- **User (U)** - Each user (or student) in the dataset is given a unique identifier. We use this feature to capture the latent ability of the user to answer the exercises.
- **User + Format (UF)** - Duolingo provides 3 formats in their dataset - `reverse_tap`, `reverse_translate` and `listen`. Each exercise can belong to one of the formats. We use a combination of user modelling and exercise format as our feature. The intuition being that the performance of a user depends on the format of the exercise.
- **Session (S)**- In the data, we find that there are 3 types of sessions - lesson, practice and test. We simply encode this information as a feature for our model.

3.2 Context Features

As mentioned in related works section, we use ideas from zone of proximal development and introduce context based features which could assist

the student in answering a particular instance in an exercise. We use these context features for all the 3 formats. [1]

- **Previous-Current Token POS and Current-Next Token POS (PCPOS, CNPOS)** - The user may implicitly learn the structure of the language. Therefore, we encode two features - Previous Token POS and Current Token POS as one of the features and Current Token POS and Next Token POS as the other feature.

- **Previous-Current and Next-Current Token Metaphone (PCM, CNM)** - We realize that sounds or phonemes can play a vital role in this task. Therefore, we make use of metaphones to represent the phonemes. Although, we use this feature in all the three tracks, we make use of English language rules to compute the metaphones in other languages as well. We encode the metaphonic combination of Previous Token and Current Token as a feature in our model. We do the same with Current Token and Next Token.

- **Previous-Current Token and Current-Next Token (PCT, CNT)** - We use the combination of Previous and Current instance token as a feature in our model. Likewise, we use a combination of Current instance token and Next instance token in the exercise as a feature.

- **First Token (FT)** - We also investigate the influence of First Token in each exercise. We normalize the First Token by lowering the case and then use it as a categorical feature.

4 Our Model

Recent works in the described in section 3 have encouraged us to use a simple logistic regression model. The equation of logistic regression is as follows:

$$P(y \mid x) = \frac{1}{1 + exp(\vec{w} \cdot \vec{f}(x, y))} \quad (1)$$

where $\vec{w}$ is the weight vector and $\vec{f}(x, y)$ is the sparse feature vector.

We use the same model in all three tracks of the competition - English-Spanish, Spanish-English and French-English.

For training the model, we make use of an L2 regularized Stochastic Gradient Descent algorithm to minimize the error, thereby maximizing the likelihood of a class. We also store feature counts to reduce the learning rate of frequently occurring features. Through trial and error, we adjusted the learning rate and prior variance for the model.

Additionally, we also experimented with Hal Daume's MegaM tool[2] through the NLTK interface. The MegaM tool looks to maximize the log likelihood of a class. Our initial results with this approach did not seem as promising as the SGD based logistic regression model. Therefore, we decided to proceed with the former.

5 Evaluation

We experiment with the features mentioned in the section 3 and evaluate the model on the development data of all three languages. Our results in all three languages were promising which encouraged us to make use of the same features with the Test set as well.

5.1 Development

The results for our model with the development set can be found in Table 1. Our results consistently indicate that a context based approach for language acquisition modeling gives good performance.

5.2 Test

We use the development data as part of our training data while evaluating our model on the test data. The results are found in Table 2. Our model beats the Duolingo's baseline model by a good margin in all three language tracks.

We note that our baseline model with all the context features gives best AUROC scores in two tracks. However, there is small dip in the AUROC in French-English track. As a future work, it would be interesting to investigate further into this decrease in performance.

[1] Our experiments with the development set indicated that Current-Next Token, Current-Next Token POS and Current-Next Token Metaphone feature reduced the AUROC when the format was listen. Therefore, in our model, we consider the above mentioned features only when the format is reverse_tap or reverse_translate.

[2] http://legacydirs.umiacs.umd.edu/~hal/megam/version0_3/

Model	en_es	es_en	fr_en
Duolingo's Baseline	0.773	0.746	0.771
Baseline	0.782	0.754	0.779
Baseline + (PCPOS, CNPOS)	0.801	0.776	0.794
Baseline + (PCPOS, CNPOS) + (PCM, CNM)	0.816	0.791	0.811
Baseline + (PCT, CNT) + (PCPOS, CNPOS) + (PCM, CNM)	0.820	0.792	**0.813**
Baseline + (PCT, CNT) + (PCPOS, CNPOS) + (PCM, CNM) + FT	**0.820**	**0.792**	0.812

Table 1: AUROC scores for our model in different language tracks on the development dataset

Model	en_es	es_en	fr_en
Duolingo's Baseline	0.774	0.746	0.771
Baseline + (PCPOS, CNPOS) + (PCM, CNM)	0.817	0.788	0.810
Baseline + (PCT, CNT) + (PCPOS, CNPOS) + (PCM, CNM)	0.821	0.789	**0.812**
Baseline + (PCT, CNT) + (PCPOS, CNPOS) + (PCM, CNM) + FT	**0.821**	**0.790**	0.811

Table 2: AUROC scores for our model in different language tracks on the test dataset

6 Future Work

Recent works in language acquisition through Code-Mixed text have suggested that providing favorable textual context for learners can be an effective strategy. We suggest that a similar strategy would be useful in the Duolingo Application. We would like to extend this line of thought to text readability and text simplification. It would be interesting to see if text simplification techniques could simplify sentences with an intention of assisting language learners to acquire new vocabulary while balancing out the readability of the text.

In our work we show that sound based features can play a vital role while learning. We use metaphones in our work to encode sound features in our model. We would like see if a more expressive method for encoding sound can be used to improve the model's performance. The data does not provide the translation of tokens in the user's native language. By computing the machine translation of these tokens, one could check the effect of cognateness of the word while answering the exercise.

7 Conclusion

In this paper, we show that a simple linear model with context based features gives good performance in modeling language acquisition. In our work, we conduct the feature ablation study and thoroughly evaluate the effect of these context based features in this task. Additionally, we also give direction for future work in text simplification and readability.

Code

To facilitate research and reconstruction of our approach, we have publicly released our code: https://github.com/iampuntre/slam18

References

Hao Cen, Kenneth Koedinger, and Brian Junker. 2008. Comparing two irt models for conjunctive skills. *Intelligent Tutoring Systems Lecture Notes in Computer Science*, page 796798.

H Ebbinghaus. 1885. *Memory: A Contribution to Experimental Psychology.* Teachers College, Columbia University, New York, NY, USA.

Rebecca Knowles, Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2016. Analyzing learner understanding of novel l2 vocabulary. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Philipp Koehn, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, and et al. 2007. Moses. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL 07*.

Igor Labutov and Hod Lipson. 2014. Generating code-switched text for lexical learning. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

S Leitner. 1972. *So lernt man lernen. Angewandte Lernpsychologie ein Weg zum Erfolg. Verlag*. Verlag Herder, Freiburg im Breisgau, Germany.

Nihal V Nayak, Tanmay Chinchore, Aishwarya Hanumanth Rao, Shane Michael Martin, Sagar Nagaraj Simha, GM Lingaraju, and HS Jamadagni. 2017. V for vocab: An intelligent flashcard application. In *Proceedings of ACL 2017, Student Research Workshop*, pages 24–29.

Henry L. Roediger and Jeffrey D. Karpicke. 2006. Test-enhanced learning. *Psychological Science*, 17(3):249255.

B. Settles, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani. 2018. Second language acquisition modeling. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. ACL.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

L.S. Vygotsky and M. Cole. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.