View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

Adaptive Forgetting Curves for Spaced Repetition Language Learning

Ahmed Zaidi, Andrew Caines, Russell Moore, Paula Buttery, and Andrew Rice

ALTA Institute & Department of Computer Science and Technology
University of Cambridge
15 JJ Thomson Ave
Cambridge, CB3 0FD
United Kingdom
{ahz22,apc38,rjm49,pjb48,acr31}@cam.ac.uk

Abstract. The *forgetting curve* has been extensively explored by psychologists, educationalists and cognitive scientists alike. In the context of Intelligent Tutoring Systems, modelling the forgetting curve for each user and knowledge component (e.g. vocabulary word) should enable us to develop optimal revision strategies that counteract memory decay and ensure long-term retention. In this study we explore a variety of forgetting curve models incorporating psychological and linguistic features, and we use these models to predict the probability of word recall by learners of English as a second language. We evaluate the impact of the models and their features using data from an online vocabulary teaching platform and find that word complexity is a highly informative feature which may be successfully learned by a neural network model.

Keywords: spaced repetition · language learning · forgetting curve · neural networks · adaptive learning.

1 Introduction

Optimal human learning techniques have been extensively studied by researchers in psychology [4] and computer science [16, 20, 8, 19]. The impact of learning techniques can be measured by how they affect the long-term retention of the learning materials. Measuring retention requires a model of the human forgetting curve, which plots the probability of recall over time. The first version of the forgetting curve was defined by Ebbinghaus [5] but has since been developed further by many researchers who have incorporated additional psychologically grounded variations to the model [17, 13, 9, 3, 14]. The ideal forgetting curve should adapt to learning materials as well as user meta-features (including current ability). In this study we examine the task of vocabulary learning. We investigate a range of linguistically motivated features, meta-features, and a variety of models in order to predict the probability a given learner will correctly recall a particular word.

2 Method

We use the Duolingo spaced repetition dataset [15] in order to train and evaluate our features and variety of models. The dataset is filtered for English language learners which results in approximately 4.28 million learner-word datapoints. Our models are a modification of the half-life regression model proposed by Settles & Meeder [16].

2.1 Half-Life Regression (HLR)

The half-life regression model is defined as follows:

$$p = 2^{-\Delta/h} \tag{1}$$

where p is the probability of recall, Δ is the time since last seen (days) and h is the *half-life* or strength of the learner's memory. We denote the estimated half-life by $\hat{h}_\Theta$, and it is defined as:

$$\hat{h}_\Theta = 2^{\Theta \cdot \mathbf{x}} \tag{2}$$

where Θ is a vector of weights for the features $\mathbf{x}$. The features of the model are made up of lexeme tags, one tag for each word in the vocabulary (e.g. the lexeme tag for word *camera* is *camera.N.SG*). The aim of these features is to capture the inherent difficulty of the word.

The HLR model is trained using the following loss function:

$$\ell(\mathbf{x}; \Theta) = (p - \hat{p}_\Theta)^2 + (h - \hat{h}_\Theta)^2 + \lambda ||\Theta||_2^2 \tag{3}$$

In practice, it was found that optimising for both p and h in the loss function improved the model. The true value of h is defined as $h = \frac{-\Delta}{log(p)}$. p and $\hat{p_\Theta}$ are the true probability and model estimated probability of recall, respectively.

2.2 HLR with Linguistic/Psychological Features (HLR+)

We now expand on the HLR model by adding additional linguistic, psychological and meta-features to $\mathbf{x}$. We refer to this model as HLR+. The features include *word complexity* scores estimated by a pre-trained model [6], *mean concreteness* scores and *percent known* based on human judgements [2], *SUBTLEX* word frequencies [18] and *user ids*.

The motivation for including complexity as a feature is based on the intuition that the more complex the word, the harder it is to remember. Concreteness is included based on previous work showing that concrete words are easier to remember than abstract words because they activate perceptual memory codes in addition to verbal codes [10]. SUBTLEX is the relative frequency of an English word based on a corpus of 201.3 million words: we hypothesise that more frequent words are more likely to be encountered and reinforced during the time since last

seen Δ. Similarly, we expect that 'percent known' (the proportion of respondents familiar with each word based on survey data) will correlate with probability of recall. Lastly, we include user id to capture latent behavioural aspects about the learners.

2.3 Complexity-based Half-Life Regression (C-HLR+)

In addition to adding new features, we now describe a new model that modifies the p such that it directly incorporates word complexity. Gooding et al. [6] derived *word complexity* to express perceived difficulty. We hypothesise that this will correlate with probability of recall. As the complexity of the word rises, the forgetting curve will become steeper. Therefore, the new model is as follows:

$$p = 2^{-\Delta \cdot C_i / h} \tag{4}$$

where C is the mean complexity for word i. We define estimated half-life $\hat{h}_\Theta$ as $2^{\Theta \cdot \mathbf{x}}$ where $\mathbf{x}$ is a vector composed of all of the features described in Section 2.2.

2.4 Neural Half-Life Regression (N-HLR+)

Motivated by the recent success of neural networks, we now describe the N-HLR+ model which replaces $\hat{h}_\Theta = 2^{\Theta \cdot \mathbf{x}}$ with a neural network. The network can be described as follows:

$$\hat{h_\Theta} = ReLU(\mathbf{x} \cdot \mathbf{w_1}) \cdot \mathbf{w_2} \tag{5}$$

where the network contains a single hidden layer. $\mathbf{x}$ is a vector of input features, $\mathbf{w_1}$ is the weight matrix between the inputs and the hidden layer and $\mathbf{w_2}$ is the weight matrix between the hidden layer and the output. We use the same loss function as HLR which optimises for both p and h.

2.5 Evaluation and Implementation

We use mean absolute error (MAE) of probability of recall for a lexical item as our evaluation metric which, despite some known problems [11], is in line with previous work [16]. MAE is defined as: $\frac{1}{D}\sum_{D}^{i=1} |p - \hat{p_\Theta}|_i$, where D is the total data instances.

We divided the Duolingo English data into 90% training and 10% test. We trained all non-neural models (e.g. HLR, HLR+, C-HLR) using the following parameters which were tuned on the first 500k data points — learning rate: 0.001, alpha α: 0.01, λ: 0.1. For all neural models (e.g. N-HLR), we used — learning rate: 0.001, epochs: 200, hidden dim: 4.

Table 1. Evaluation of forgetting curve models. Pimsleur and Leitner are previous methods of modelling the forgetting curve.

Model	MAE↓
Pimsleur[12]	0.396
Leitner[7]	0.214
Logistic Regression	0.196
HLR[16]	0.195
HLR-lex[16]	0.130

Model	MAE↓
HLR+	0.129
C-HLR+	0.109
N-HLR+	**0.105**
CN-HLR+	**0.105**

3 Results and Discussion

We can see in Table 1 that HLR+ did not perform much better than HLR. By modifying the loss function to include complexity as a parameter in the C-HLR+ model, we considerably improved the performance of our model. This was in line with our hypothesis that more complex words are forgotten faster and thus are an important feature in modelling the forgetting curve.

The N-HLR+ model provided additional improvements to the C-HLR+ model. This is due to the fact that neural models are better at capturing non-linearities between the features and the expected output. Furthermore, when compared to the N-HLR+ model we can see that including complexity into the loss function (CN-HLR+) provides no clear improvements in performance. This is because the model learns to place more importance on the *complexity feature*. We confirm this by analysing the average weights in the hidden layer of the model. The model learns to give greater importance to word complexity, percent known, and concreteness respectively. It does not however, learn much from the user id and SUBTLEX. This is probably due to the fact that a single dimension for capturing user behaviour is not sufficient and that SUBTLEX does not adequately represent learners' experience with English as a second language.

4 Conclusion

We present a new model for adaptively learning a forgetting curve for language learning using a modified HLR loss function and a neural network. We incorporate linguistically and psychologically motivated features and show that word complexity is an important feature in predicting probability of recall for a vocabulary item. Furthermore, we illustrate that neural networks can capture the importance of word complexity while a simple HLR fails to take advantage of that signal. This work lays the foundation for work in neural approaches to understanding language learning over time. Future work in this area includes incorporating high-dimensional user embeddings to capture user specific signals that might influence the forgetting curve, and also different models such as Pareto and power functions which have been proposed in prior work [1].

Acknowledgements

This paper reports on research supported by Cambridge Assessment, University of Cambridge.

References

1. Averell, L., Heathcote, A.: The form of the forgetting curve and the fate of memories. Journal of Mathematical Psychology **55**, 25–35 (2011)
2. Brysbaert, M., Warriner, A.B., Kuperman, V.: Concreteness ratings for 40 thousand generally known English word lemmas. Behavior research methods **46**(3), 904–911 (2014)
3. Choffin, B., Popineau, F., Bourda, Y., Vie, J.: DAS3H: modeling student learning and forgetting for optimally scheduling distributed practice of skills. In: Proceedings of The 12th International Conference on Educational Data Mining (EDM) (2019)
4. Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., Willingham, D.T.: Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest **14**(1), 4–58 (2013)
5. Ebbinghaus, H.: Ueber das gedächtnis (1885)
6. Gooding, S., Kochmar, E.: Complex word identification as a sequence labelling task. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1148–1153 (2019)
7. Leitner, S.: So lernt man lernen: angewandte Lernpsychologie–ein Weg zum Erfolg. Herder. (1972)
8. Moore, R., Caines, A., Elliott, M., Zaidi, A., Rice, A., Buttery, P.: Skills embeddings: a neural approach to multicomponent representations of students and tasks. In: Proceedings of The 12th International Conference on Educational Data Mining (EDM). vol. 360, p. 365. ERIC
9. Mozer, M.C., Wiseheart, M., Novikoff, T.P.: Artificial intelligence to support human instruction. Proceedings of the National Academy of Sciences **116**(10), 3953–3955 (2019)
10. Paivio, A.: Imagery and verbal processes. Psychology Press (2013)
11. Pelánek, R.: Metrics for evaluation of student models. Journal of Educational Data Mining **7**(2), 1–19 (2015)
12. Pimsleur, P.: A memory schedule. The Modern Language Journal **51**(2), 73–75 (1967)
13. Reddy, S., Levine, S., Dragan, A.: Accelerating human learning with deep reinforcement learning. In: NeurIPS workshop: teaching machines, robots, and humans (2017)
14. Rubin, D.C., Wenzel, A.E.: One hundred years of forgetting: A quantitative description of retention. Psychological Review **103**(4), 734 (1996)
15. Settles, B.: Replication Data for: A Trainable Spaced Repetition Model for Language Learning (2017). https://doi.org/10.7910/DVN/N8XJME, https://doi.org/10.7910/DVN/N8XJME
16. Settles, B., Meeder, B.: A trainable spaced repetition model for language learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1848–1858 (2016)

17. Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., Gomez-Rodriguez, M.: Enhancing human learning via spaced repetition optimization. Proceedings of the National Academy of Sciences **116**(10), 3988–3993 (2019)
18. Van Heuven, W.J., Mandera, P., Keuleers, E., Brysbaert, M.: SUBTLEX-UK: A new and improved word frequency database for British English. The Quarterly Journal of Experimental Psychology **67**(6), 1176–1190 (2014)
19. Zaidi, A.H., Caines, A., Davis, C., Moore, R., Buttery, P., Rice, A.: Accurate modelling of language learning tasks and students using representations of grammatical proficiency. In: Proceedings of The 12th International Conference on Educational Data Mining (EDM) (2019)
20. Zaidi, A.H., Moore, R., Briscoe, T.: Curriculum Q-learning for visual vocabulary acquisition. In: Proceedings of Visually-Grounded Interaction and Language (ViGIL), NeurIPS (2017)