Curriculum Q-Learning for Visual Vocabulary Acquisition

Ahmed H. Zaidi[1,2], **Russell Moore**[1], **and Ted Briscoe**[1]

[1]Computer Laboratory, University of Cambridge
[2]Catalyst AI Ltd., UK
{ahmed.zaidi,rjm49,ted.briscoe}@cl.cam.ac.uk

Abstract

The structure of curriculum plays a vital role in our learning process, both as children and adults. Presenting material in ascending order of difficulty that also exploits prior knowledge can have a significant impact on the rate of learning. However, the notion of difficulty and prior knowledge differs from person to person. Motivated by the need for a personalised curriculum, we present a novel method of curriculum learning for vocabulary words in the form of visual prompts. We employ a reinforcement learning model grounded in pedagogical theories that emulates the actions of a tutor. We simulate three students with different levels of vocabulary knowledge in order to evaluate how well our model adapts to the environment. The results of the simulation reveal that through interaction, the model is able to identify areas of weakness, as well as push students to the edge of their *zone of proximal development*. We hypothesise that these methods can also be effective in training agents to learn language representations in a simulated environment where it has previously been shown that order of words and prior knowledge play an important role in the efficiency of language learning.

1 Introduction

With the rise of machine learning and tasks such as automated teaching and assessment, there is an increased interest in understanding how machine learning models can be grounded in theories of language acquisition. Additionally, with an abundance of learner data in archive and generation, we now have an avenue through which we can not only evaluate our theories of learning, but also explore whether these theories can be used to train agents for the purpose of general AI.

Language Acquisition is a multidisciplinary field that overlaps with linguistics, psychology, neuroscience, philosophy, and more recently computer science. At the intersection of language acquisition and pedagogy lie theories of educational practices for language learners, including for example, an optimal curriculum for both L1 and L2 learners. A *curriculum* is a guide that helps teachers decide what content to present and the order of which it needs to be presented. The aim of a curriculum is to provide a highly structured method of introducing concepts in order to maximise the rate of learning.

The idea of a curriculum to facilitate the rate of learning has been discussed from the perspective of *animal training* [1, 2], where it is defined as *shaping*. It has also been referenced in an educational framework [3] where the author introduces the idea of a *spiral curriculum*, a process by which complex information is first presented in a simplified manner and then revisited at a more difficult level later on. Similarly Vygotsky, from the view of language acquisition, introduces the idea of *scaffolding* in order to provide contextual support for more complex ideas using simplified language or visuals. Elman [4] draws parallels between the effectiveness of staged learning in humans, and in artificial neural models. All of these concepts have been discussed in different fields but reference

Visually Grounded Interaction and Language Workshop at NIPS 2017, Long Beach, CA, USA.

the same underlying idea of presenting information in a structured manner in order to exploit prior knowledge.

Bruner [5] argues that the role of the teacher is not to present information by rote learning but rather facilitate the learning process in order to teach students to become *active learners*: put simply, they are "learning to learn". There are many factors that teachers need to consider when constructing a curriculum to achieve this goal, namely the *difficulty* and *appropriateness* of content.

Difficulty is measured relative to the *zone of proximal development (ZDP)*, introduced by Vygotsky, which is a representation of what a learner is capable of achieving without help, with *some* help, and of concepts that are beyond the learner's current ability. Appropriateness is a measure of whether content being presented is within the ZPD or, in the case of scaffolding, comprises material from within the ZPD.

Determining difficulty and appropriateness is traditionally a very laborious and resource intensive task which entails experts conducting focus groups and analysis to decide where a particular question or topic sits in the curriculum. This method is not only inefficient, it also assumes a static curriculum for all students.

To address these limitations, we propose the use of reinforcement learning (RL) in order to learn an optimal policy and curriculum for each student for the task of visual vocabulary acquisition. Through this, we also discuss the similarities between the properties and features of RL and those of language acquisition. We evaluate our models by simulating three types of student at different levels of proficiency *(beginner, intermediate, and advanced)*. We find that the system is able to identify the difference in proficiency and adapt its curriculum to reflect this difference.

Previous uses of RL in pedagogy include [6] where it is used to teach students arithmetic, aiming to minimise the time taken to answer questions. [7, 8] teach students database design using Q-learning. Both [6] and [7, 8] evaluate results on simulated students. [9] use RL for maths while [10] use it for physics. However, as far as we know, no previous work has been done in the space of visual lexical acquisition where the principles of RL have explicitly been related to theories of language acquisition.

The importance of curriculum learning in training deep learning models and agents has also been discussed by [11] where its use is shown to facilitate the generalisation as well as the rate of convergence and training of deep learning networks. [12] also illustrate the need for some form of curriculum to improve the rate of learning for agents in a 3D simulation. However, it is worth noting that no explicit RL is used to model curriculum by either [11] or [12].

2 Curriculum Q-Learning

In order to automate the process of curriculum learning for visual vocabulary acquisition, we must first identify the key components of our RL system. The agent in this task is the automated tutor that must learn what information to present to the student. The environment is the student with whom the agent is interacting.

We assume that the student is a learner of English who has reached a given level on the *Common European Framework of Reference* (CEFR) scale. CEFR is an international standard for describing language ability, using a six point scale, from A1 for beginners, up to C2 for those who have mastered language.

The RL algorithm used by our proposed system is Q-Learning, an *off-policy* algorithm for Temporal Difference (TD) Learning. Q-Learning can be defined as follows:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q^{\pi}(s',a') - Q(s,a)] \tag{1}$$

where $Q(s,a)$ is the Q-value of a state s and action a tuple. The α is the learning rate and γ is the discount factor. γ models the fact that future rewards are less valuable than immediate rewards at a given time t.

A policy π maps states s to actions a. The aim of the Q-Learning algorithm is to find an optimal policy π such that it maximises the long-term cumulative reward. The policy achieves this by acting greedily and taking the action that presents the maximum Q-value given the state such that $\max_{a \in A} Q^{\pi}(s_t, a)$.

In action selection, there is a trade-off between *exploiting* what you have learnt so far and *exploring* other state-action tuples. In this task we model that using ϵ-greedy. This means the policy will, for most part, select the actions that provide the highest estimated future reward given the state. However, with a probability of $1 - \epsilon$, an action will be selected randomly and independently from a uniform distribution. Action selection is usually drawn from a Q-Table which is a table that stores all state-action Q-values.

In this task, a policy can be viewed as a curriculum as it decides what should be shown and in what order. In order to learn a curriculum for vocabulary acquisition, we incorporate two models, the *CEFR level model* (See Figure 4) and *word level model* (See Figure 5). The CEFR level model has 6 states which are defined by the 6 CEFR levels. The actions are whether the student should progress to the next level, stay in the current level, or go back a level. The word level model has two states: active (show the word), inactive (hide the word). The actions are remain in the current state or toggle state. This architecture ensures that there is also an estimated long-term reward associated with showing a student a particular word.

Modelling reward is often viewed as a challenging task in RL. For this application, a student is rewarded negatively (-1) for getting a question correct and positively (+1) for getting it incorrect. The motivation behind using these values is grounded in how we learn. The RL model acts greedily and takes the action with the maximum reward, so if we review a concept we understand, then we are not gaining knowledge by reviewing it again. Thus its value should be reduced. Alternatively, if we get a question wrong, the benefit of reviewing that word is higher, and thus we should increase the associated Q-value.

To evaluate the students' understanding, we present a word in the form of an image. The objective for the students is to describe the image, and based on their response, the Q-Learning algorithm and thus the policy is updated. A valid response is defined by the target word associated with the image or a near synonym of that target word, which is automatically generated by looking at the top 10 nearest words to the target word in a pre-trained *word2vec* model [13]. The use of images was motivated by the ease of generating teaching materials and widespread use of flashcards for vocabulary learning. Additionally, there are countless studies that indicate the effectiveness of images for learning [14].

3 Experiments

For the CEFR level model, we use a learning rate α of 0.1, a discount rate γ of 0.9 and an ϵ value of 0.95. The word level model uses an α of 0.1, a γ of 0.9 and an ϵ value of 1 in order to prevent words randomly going into an inactive state.

To evaluate the performance of our system, we simulated three types of students at varying levels of proficiency: *beginner, intermediate and advanced*. In this case, we modelled the student's probability of getting a question correct as a negated Gompertz[15] distribution:

$$P(success \mid u, q) = 1 - \exp(-b \ \exp(-c(l(q) - l(u)))) \tag{2}$$

where $l(u)$ denotes the level of user u calibrated to a scale of $[0, 6]$. Each integer in the scale represents a corresponding CEFR level from A1 to C2 (*e.g.* $0 \rightarrow$ A1, $1 \rightarrow$ A2, *etc.*). $l(q)$ represents the level of an item q (i.e. a word which must be guessed from an image) calibrated to the same scale. The parameter b determines the probability of success when student and item level match. This is set to $ln(0.75)$ to model a 'typical' pass rate of 75%. The calibrated curve is shown in Appendix C. The curve is flatter at the lower end as students may be expected to be comfortable with most of the material at lower CEFR levels than their own, whereas at higher levels, their ability is more uncertain. We ran simulations where each student had 100 interactions with the system. An interaction can be defined as when a student responds to a question.

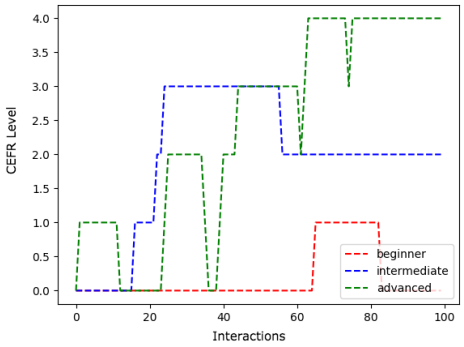

Figure 1: CEFR levels determined by the agent for students of varying levels of proficiency over 100 interactions

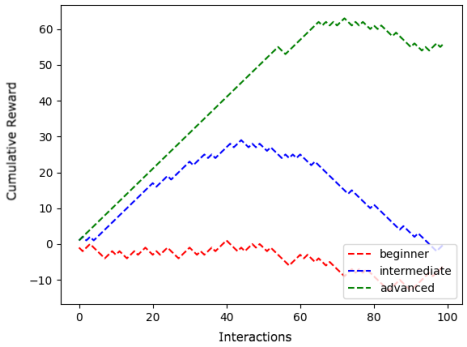

Figure 2: Cumulative reward earned by students of varying levels of proficiency from the agent over 100 interactions

3.1 Results

The results from Figure 1 show how the agent responds to the various proficiency levels. The beginner student remains relatively constant around A1 and A2 which is reflective of the student's current level. The intermediate student continually increases in CEFR level until level 3 (B2). The advanced student, although tested with material beneath the actual level of proficiency, eventually reaches an advanced or higher CEFR level. We can also see that the agent tutor pushes the student to what can be interpreted as the edge of their ZDP. Figure 2 illustrates how the cumulative reward of the students varies for students at different proficiencies. The curve experiences a downward slope as the students reach their current level of vocabulary and are now being pushed to understand more advanced material.

4 Discussion

We have shown through the use of simulations, that we can effectively model a personalised curriculum for vocabulary acquisition using Q-Learning. Figure 1 and Figure 2 show clear indications of varying agent behaviour for students at different levels of lexical proficiency. However, beyond that, we have set up a framework that can be used in the future to extrapolate the difficulty and appropriateness of new material. The system will serve as a test bed that will yield metrics to determine where the content fits in the curriculum. Although this is foundational work, it lays the building blocks for future pedagogically inspired RL architectures.

Through this work, we have also shown that there are many similarities between the principles of RL and theories of language acquisition. Specifically, parallels can be drawn between the concept of ϵ-greedy and Krashen's Input Hypothesis or the $i+1$. The Input Hypothesis states that students learn by comprehending language that is slightly above their current language level. The interactions between the agent and the environment in RL is analogous to the social interaction approach to language acquisition, specifically the equal importance of input and output. For this reason, we use the Q-Learning algorithm as opposed to the SARSA algorithm mainly due to the properties of Q-Learning that ensure an "optimal path" is followed i.e. the minimum number of steps to reach our goal (language fluency).

However, there is scope for substantial extensions in this space. Deploying the system on-line in order to collect user data will allow us to validate and improve our existing models. Incorporating memory and spaced repetition learning [16], a phenomenon initially documented by Ebbinghaus (1885), in order to optimise the policy and emulate cognitive processes is also an important extension that may have a great impact on the learning output. Using deep learning models to approximate the Q-value will allow the system to capture additional signals pertinent to language acquisition. Additionally, moving towards an adaptive reward model that reflects difficulty to encourage memory retention.

All of these models can also be applied to agents instead of students. As discussed previously, [12] indicated the need for a curriculum in order to effectively train an agent in the simulated environment. Creating a dynamic environment guided by a curriculum grounded in pedagogically inspired RL may result in improved learning rates for the agent.

Acknowledgements

We thank Wenchao Chen who helped develop the back-end of our web-based platform.

References

[1] Burrhus Frederic Skinner. Teaching machines. *Science*, 128(3330):969–977, 1958.

[2] Gail B Peterson. A day of great illumination: Bf skinner's discovery of shaping. *Journal of the Experimental Analysis of Behavior*, 82(3):317–328, 2004.

[3] Jerome S Bruner. *The process of education:[a searching discussion of school education opening new paths to learning and teaching]*. Vintage Books, 1960.

[4] Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[5] Jerome S Bruner. The act of discovery. *Harvard educational review*, 1961.

[6] Joseph Beck, Beverly Park Woolf, and Carole R Beal. Advisor: A machine learning architecture for intelligent tutor construction. *AAAI/IAAI*, 2000:552–557, 2000.

[7] Ana Iglesias, Paloma Martínez, Ricardo Aler, and Fernando Fernández. Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Applied Intelligence*, 31(1):89–106, 2009.

[8] Ana Iglesias, Paloma Martinez, and Fernando Fernández. An experience applying reinforcement learning in a web-based adaptive and intelligent educational system. 2003.

[9] Kimberly N Martin and Ivon Arroyo. Agentx: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems. In *Intelligent Tutoring Systems*, pages 564–572. Springer, 2004.

[10] Joel R Tetreault and Diane J Litman. Comparing the utility of state features in spoken dialogue using reinforcement learning. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 272–279. Association for Computational Linguistics, 2006.

[11] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[12] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojtek Czarnecki, Max Jaderberg, Denis Teplyashin, et al. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*, 2017.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] Michael P Verdi, Janet T Johnson, William A Stock, Raymond W Kulhavy, and Polly Whitman-Ahern. Organized spatial displays and texts: Effects of presentation order and display type on learning outcomes. *The Journal of Experimental Education*, 65(4):303–317, 1997.

[15] Charles P Winsor. The gompertz curve as a growth curve. *Proceedings of the national academy of sciences*, 18(1):1–8, 1932.

[16] Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *ACL (1)*, 2016.

Appendix A Curriculum Q-Learning System Overview

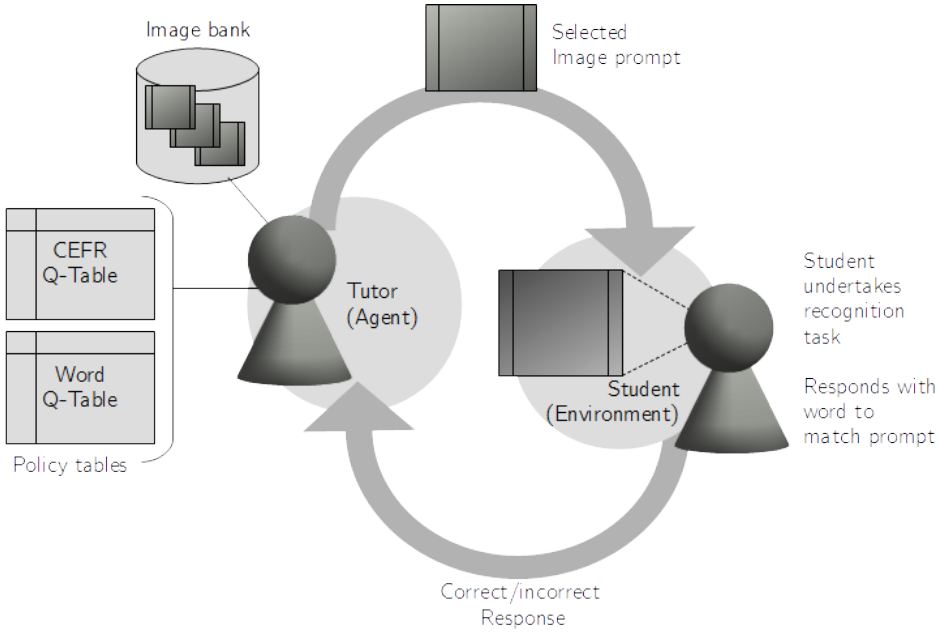

Figure 3: Overview of the system. A simulated student takes the place of a human actor in our study.

Appendix B CEFR Level and Word Level Q-Tables

CEFR	Back	Remain	Forward
A1	0	1	0
A2	0	1	0
B1	0	1	0
B2	0	1	0
C1	0	1	0
C2	0	1	0

Figure 4: Q-Table for CEFR Level model. The table is biased towards the remain state at initialisation to ensure that the student remains in the current CEFR level until it is no longer beneficial to do so.

Status	Remain	Toggle
Active	0.1	0
Inactive	0	0

Figure 5: Q-Table for Word Level model. The table is biased towards the remain state at initialisation to ensure that the word is always seen at least once for each student.

Appendix C Negated Gompertz Curve

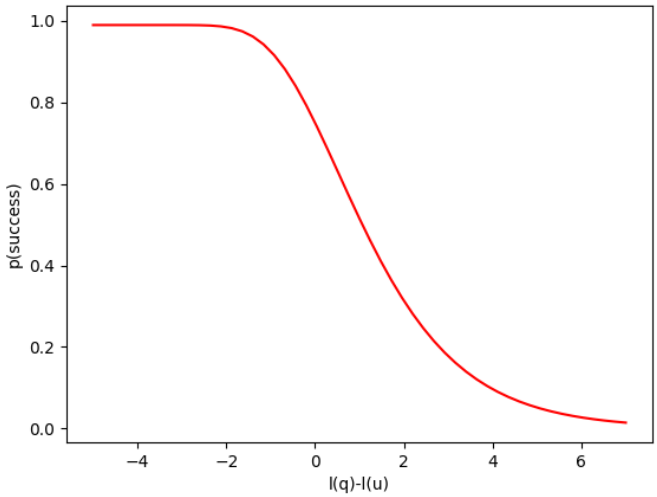

Figure 6: Gompertz curve used as a model to simulate student success probabilities.

Appendix D Preview of Web-based Curriculum Q-Learning

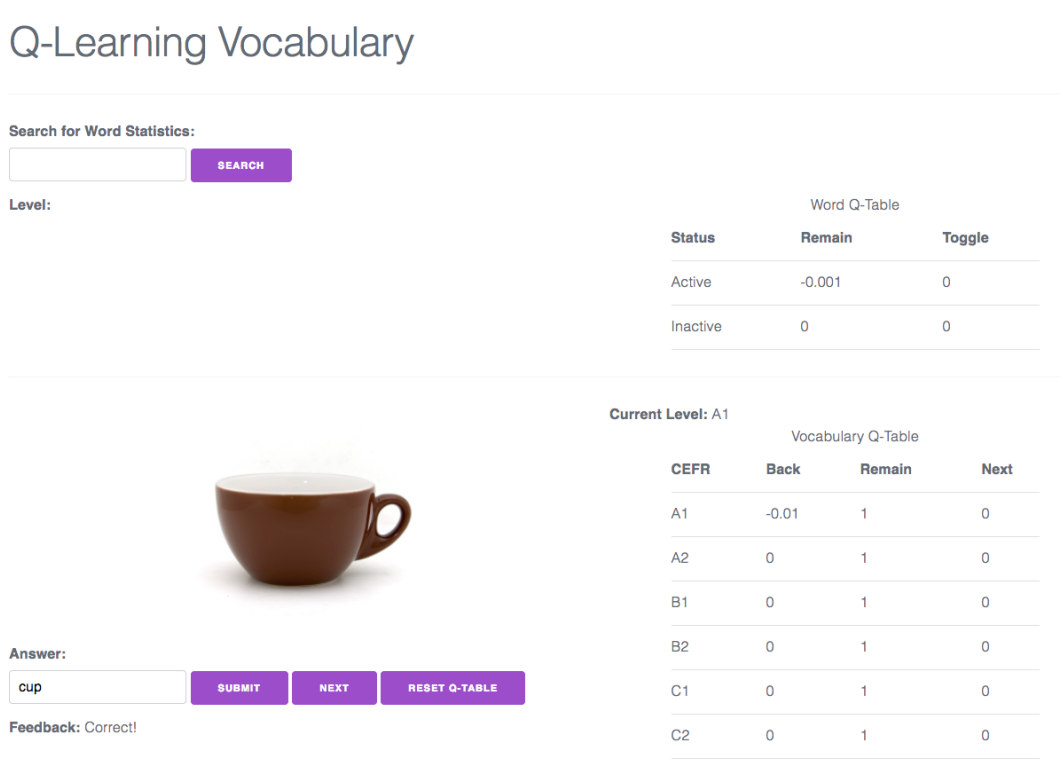

Figure 7: A preview of the web-based Curriculum Q-Learning platform.