Verbivore: Learnersourcing Vocabulary Flashcards

Paul Grau*
KAIST
grau@kaist.ac.kr

Oisin Daly Kiaer*
KAIST
oidaki@kaist.ac.kr

Yoo Jin Lim*
KAIST
yjlim@kaist.ac.kr

ABSTRACT

Learning vocabulary through standard flashcards yields poor results, due to the individuals' lapses in motivation and inspiration in the face of tedious work, along with a lack of good guidelines to select what and how to learn. We present Verbivore, a novel vocabulary flashcard studying and -generation application that leverages crowd-work to iteratively improve its set of cards and attempts to lessen the tedious busywork associated with flashcard creation. Users study the cards and are periodically confronted with microtasks to help them learn. The output of these microtasks is then leveraged to improve the cards themselves for the benefit of future learners. In a preliminary test deployment we gathered encouraging feedback while also realizing several limitations, including the importance of rich feedback for user tasks.

Author Keywords

Learnersourcing; integrated workflow; vocabulary.

INTRODUCTION

Learning vocabulary is a fundamental activity that affects a wide range of academic skills. Several of life's milestones additionally demand some level of vocabulary knowledge; standardized tests with the American SAT or GRE being prime examples. For these reasons, many people are highly motivated to improve their vocabulary, and attempt to do so in various ways. Hand-made flashcards are one way, flashcard study applications are another. Common for all major flashcard studying apps is that their content is to a very large degree expert-generated, or made by individual amateurs. In this paper we present Verbivore, an alternative flashcard studying application that uniquely crowdsources its content at a very granular level, having learners do work related to the words that they are studying, and channeling that work back into the app to improve the content for future learners. The result is a set of content that continually and dynamically improves, which we hope will lead to a higher eventual quality than what the alternatives provide. A big part of this type of app is of course the user feedback aspect that it holds over traditional methods. Practice exercises with immediate feedback are immensely helpful for learning; they demand knowledge for the learner to proceed, mediating lapsing attention, and simultaneously making the knowledge 'stickier'. We want to achieve a similar mechanic with the work that we have our users do. Verbivore tries to stimulate its users in the same way as quizzes, but with the restriction that all its tasks must be either content-generative or -filtering in some way.

*All authors have equal contribution in this work

BACKGROUND AND RELATED WORK

Mobile applications for vocabulary learning

Recent advances in mobile computing have shown potential to alleviate issues with vocabulary learning. Since the mid-1990s, research in Mobile-Assisted Language Learning [3] has focused on using mobile systems to help students. A number of projects have tried to employ findings in learning theory (e.g. spacing effect) and motivational effects (e.g. gamification), but few have been able to sustain engagement.

There exist numerous commercial mobile applications for vocabulary learning, such as Memrise[1] and Duolingo [4]. Memrise provides libraries of flashcard decks generated by individuals for various subjects. Duolingo provides language education with gamification techniques while crowdsourcing translations. Verbivore is similar to these services in terms of core flashcard functionality. However, Memrise and Duolingo are not flexible enough to meet individual learners' needs. Verbivore allows learners to engage with the card contents at feature level, which includes definitions, example sentences, and images.

Learnersourcing

Learnersourcing is a conceptual framework in which learners collectively generate useful contents for future learners while engaging in a meaningful learning experience themselves [5]. Learners are in some ways experts and often reveal themselves to be better than experts. They have experienced the difficulties of learning and thus can be guided to generate high-quality learning contents for others.

One example of learnersourcing application is ASL-Flash [2], a system for learning American Sign Language (ASL) that includes an ASL dictionary and a flashcard learning tool. This work shows a flashcard tool where learners' data can be put to use for a greater purpose like a dictionary. Inspired by ASL-Flash, we introduce Verbivore with a great practice to think deeply about the vocabulary they just learned.

Another example is AXIS (Adaptive eXplanation Improvement System), which dynamically improves explanations over

[1] http://www.memrise.com/

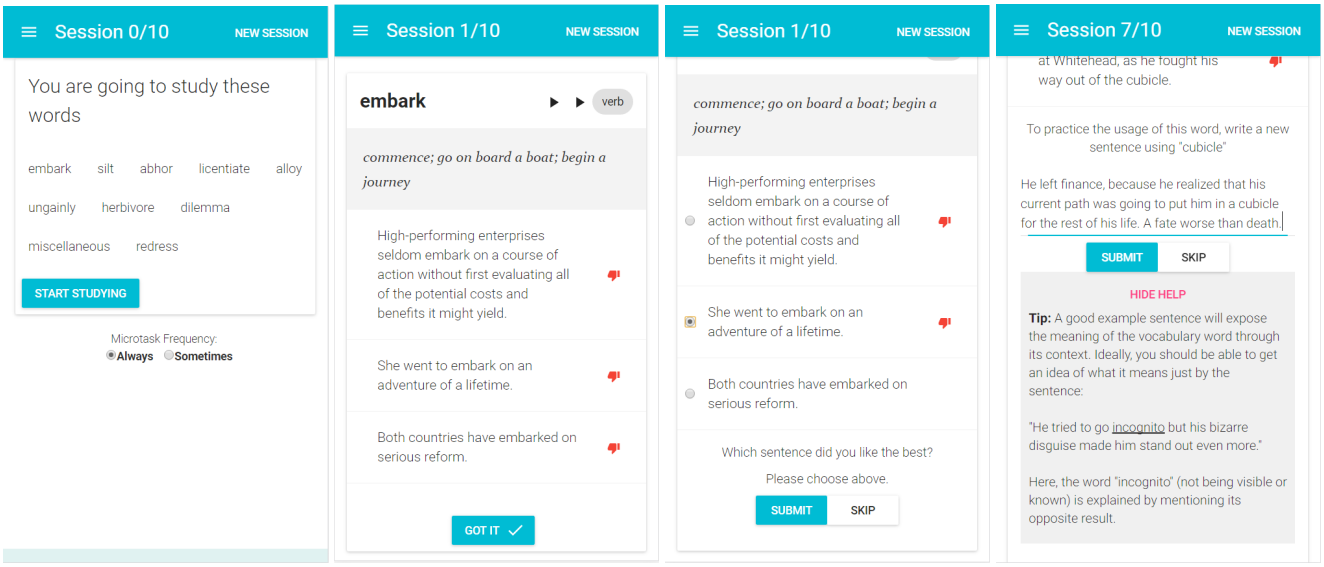

Figure 1. Main screens of Verbivore. 1. start of a session, 2. flashcard with audio clips, part of speech, definition and examples, 3. Vote task, 4. Fix task

time as a byproduct of learners' collective interactions with the content [6]. Analogous to the explanations in math problems, example sentences provide understanding to vocabulary learners, especially in flashcard learning tools. Thus, Verbivore engages learners in microtasks to dynamically improve the quality of example sentences over time.

SYSTEM

For the prototype, we developed a web application that is functional both on desktop and mobile devices. The application consists of a frontend using React and Material UI, and a backend written in Python using Flask and SQLAlchemy. We identified the following major tasks and components.

Study sessions

Learners study new words in sessions of 10 items. At the start of a new session, 10 flashcards are selected and previewed as a list. Users then go through each card one by one.

The main purpose of sessions is to give learners a short-term goal. During the session, the progress is shown as "n of 10". This is done to motivate users to continue and finish a good number of cards.

Currently, the selection of flashcards is purely random. In the future, we plan to use a number of heuristics to make the selection smarter. Words that the user already learned should be shown less. After gathering data about difficulty, we can sample a balanced set of easy and hard cards (or, according to user preference). Furthermore, we can leverage the card selection to steer users towards cards that the system hasn't collected much user data on yet.

Flashcards

The integral unit of Verbivore is a flashcard. A card contains a word and a number of user-generated features such as definitions, example sentences, and pronunciation clips. Learners view a card's contents to study a word.

Individual features can be flagged by a user for bad quality. Currently, this is only possible for example sentences. To flag, users click a small red thumbs-down button next to a sentence. They are then asked to provide a reason for the flag, which has to be one of "Wrong usage", "Bad grammar", or "Unclear".

Features to be shown in each feature category are selected using a set of heuristics explained later.

Microtasks

The main idea of Verbivore is to have learners perform small tasks while they are studying. In the current implementation, a task can be shown every time a user is done studying one flashcard, indicated by clicking "Got it" at the bottom of the card. There is a setting to either show tasks always or in 75% of cards.

There are currently three microtasks that roughly implement a Create-Fix-Verify scheme (c.f. Find-Fix-Verify [1]).

Create sentence. Learners are asked to apply the newly learned vocabulary by coming up with a new example sentence that uses that word. For novice users, this can be hard task, so the system provides a small guide to teach how to create good examples. User-provided content is automatically verified by a set of rules: the sentence must contain at least four words, of which one has to be the word in question.

Fix sentence. This task presents a sentence that was flagged by another user as being of bad quality. It asks the learners to provide another, better version of the example. When saving, the database also holds a reference to the original version, so that a graph is formed that can be utilized in the aggregation of contributions.

Vote on sentence. In the simplest of the three tasks, learners are asked to select one of the currently shown example sentences they personally like most. Votes are saved both as an

individual entry in the database as well as a total number of votes for each feature.

Task selection. Every time a user is supposed to work on a microtask, the system randomly selects between these three tasks. The distribution during the deployment was 30% *Create*, 40% *Fix* and 30% *Vote*, but for cards that did not have any flagged sentences and thus nothing to fix, it fell back on *Vote*. Ideally, this distribution should be adjustable during the deployment to react to usage trends and feedback.

From early feedback we learned that learners sometimes do not feel comfortable in contributing to words that they just learned and barely understood. That is why all of the tasks mentioned above include a "Skip" button to eliminate this frustration. The disadvantage of making microtasks non-mandatory is a lower expected rate of user contributions.

Aggregation and quality control

These three (or four, including the flagging system) tasks form a loop to gradually increase the quality of a card's features. Learners create and iterate on example sentences; voting is used to signify good quality while flagging is used to mark bad-quality contributions.

On each card, learners are presented with a small selection of all features in the database. We initially envisioned using a variant of a multi-armed bandit and Thomson sampling like in [6], but settled for a more simple approach for this initial version. According to the votes, each sentence is assigned a weight which forms a distribution by which a number of sentences to be shown are sampled. Currently, each card shows three example sentences. The goal is to show a diverse set of examples, biased towards higher-quality ones but also showing new or fixed sentences to some users.

Sentences that received a number of flags are automatically excluded (at the moment, the threshold is 3, but this number is yet to be empirically verified).

EVALUATION

We advertised the application through our personal social networks. In just a number of days, we received a lot of useful quantitative and qualitative feedback that we would like to present here.

To gather data on how our testers use the app, we relied on three methods. Internally, we recorded session progress and contributions. Using Google Analytics, we collected more general data on visitors such as geographic location, and also time-based data (page views and certain events) to be able to reconstruct each user's interaction. Lastly, we asked testers to complete a survey after they finished one study session.

For our test deployment, we seeded the database with a selection of 50 words taken from a list of Graduate Record Examinations (GRE) vocabulary and gathered definitions and examples from the Pearson Dictionaries API[2]. We chose a small number of words so that our limited number of testers got to interact with other users' contributions on the same cards.

Users		40
Study sessions	total	48
	partly completed[3]	31 (64.6%)
	completed	19 (39.6%)
Median session duration		9:35
Cards viewed		232
Sentence contributions	votes	125
	flags	24
	improved	5
	created	34

Table 1. Quantitative results of initial 48 hours of deployment

For some overall usage statistics, see Table 1. According to Google Analytics, our testers came from a wide range of countries (ca. 60% from South Korea, 10% from each Denmark and the US, and a small number of users from Germany, United Kingdom and Ireland). Circa 1/3 of them used a mobile browser to access the application.

Even though the number of participants is too low to draw any final conclusions, the general feedback we received was very encouraging. Many users commented that they liked the idea of doing various kinds of small tasks while studying. 15 participants (60%) said it certainly helped them learn the words. Most also agreed with the idea of sessions of 10 flashcards (14, 56%). Upon inspection, most of our participants' contributions were of high quality. One user's creations received a total of 10 votes from other learners.

Nevertheless, it seems that our design did not convey the quality and advantages of the crowdsourced material well enough. 13 participants (56.5%) said it is not a good idea to have learners generate the cards' contents. They preferred if a credible (expert) source would provide the work. While the current design did not highlight any benefits of the user-generated content, some participants identified them themselves. They commented that sentences created by other learners can be humorous or more natural and thus easier to understand than examples from a dictionary which are usually sourced from literature or newspaper articles.

So, it seems that while the given work helps users to learn the material, they are not convinced of the validity of the learnersourced content, the prevalent sentiment among the qualitative comments being a lack of trust that learners can make proper content, signaling issues with quality control.

See Table 2 for a complete report on the survey results.

We also received a lot of constructive feedback to improve the design of the cards and workflows. We will address these and some other limitations in the following section.

[2] http://developer.pearson.com/apis/dictionaries

[3] completed at least one flashcard

Do you feel that the microtasks helped you learn the words?	
Eh, not really.	3
Yeah a bit, I guess...	5
Yes, certainly!	15
(No opinion)	2
Do you think having people work together is a good way to generate the cards?	
It's not much of a difference one way or the other.	2
No, I think having learners make the content is a bad idea.	13
No, but for another reason	1
Yes, but for another reason	1
Yes, it's great! It makes people feel involved and yields high quality cards.	6
(No opinion)	2
How was the session length for you?	
A little too long. It got pretty tedious near the end.	6
I don't think there should be sessions at all. I should just be able to quit whenever.	2
Just right. 10 is my favorite number!	14
Keep 'em coming! (>15 cards per session)	1
(No opinion)	2

Table 2. Results from the post-test survey (N=25)

DISCUSSION

On balance, we believe ourselves to have been somewhat overzealous in our ambition for the system, considering the time limitation of a class project. We set out to examine whether a crowdsourced approach to flashcard generation (and more generally, learning material creation) would be in some way preferable to the expert or single-learner approach taken by language studying apps and traditional flashcard studying respectively. As it stands, the Verbivore application cannot answer those questions confidently, on account of its feature-poverty relative to its points of comparison. Verbivore still has some issues that make it less usable than its competitors.

First, the work it gives users has very little flow to it and is not very engaging. For a "micro" task, the task of creating a sentence is rather strenuous. In its current form, it is neither fun nor motivating.

Second, we did not give our users much in the way of feedback on their interactions. We believe this to be the largest failing of the app, and the issue that we would want to fix first. Many users called for an ability to track their contributions to- and progress through the set of cards. This concept was in our original plan as well, as part of our gamification suite, but we decided to focus on the fundamentals: cards and microtasks. The result is that Verbivore is somewhat unsatisfying to use, because the users do not know how well they are performing. We believe that more involved feedback and gamification aspects can add motivation both to keep studying and to participate in the microtasks.

Along the lines of user feedback, quizzes are inherently more satisfying for the users, but they have the issue of not being very useful for the iterative card improvement that we are trying to achieve, on account of not outputting any content. They would however have been interesting to have, if not only for being able to test our users' retention after using the app. The only measure of the mnemonic value of Verbivore we have so far is the self-reported one from the survey.

Additional features that we are planning for the future include more various feature types apart from example sentences (pronunciation sound clips, images, synonyms, antonyms, connotations etc.). Another idea is to highlight the aspect of community-sourcing on the cards by showing which user contributed which part or how and when they interacted with the card, possibly in real-time. We believe that this would make the process more engaging and fun. Furthermore, we have some ideas to improve the aggregation quality. For example, we plan to take into account the graph of fixed sentences to only show one version at a time, and also use similarity measures to avoid showing too similar sentences.

One should be hesitant to cherry-pick from one's quantitative data, but we think that the concept holds merit, compared to its expert-generated-content cousins. From the promising qualitative comments we got, and from conversations with our test users, we saw some really promising signals for the learnersourced approach. To users with no preconceptions, Verbivore does not sell the concept well enough in its current form, but the idea hit home with many of our test users after a more detailed explanation.

REFERENCES

1. Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94.
2. Danielle Bragg, Kyle Rector, and Richard E Ladner. 2015. A User-Powered American Sign Language Dictionary. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1837–1848.
3. Jack Burston. 2013. Mobile-assisted language learning: A selected annotated bibliography of implementation studies 1994–2012. *Language Learning & Technology* 17, 3 (2013), 157–224.
4. Luis von Ahn. 2013. Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 1–2.
5. Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 405–416.
6. Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 379–388.