A Reading Environment for Learners of Chinese as a Foreign Language

John Lee, Chun Yin Lam, Shu Jiang
Department of Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk, mickey1224@gmail.com, jshmjs45@gmail.com

Abstract

We present a mobile app that provides a reading environment for learners of Chinese as a foreign language. The app includes a text database that offers over 500K articles from Chinese Wikipedia. These articles have been word-segmented; each word is linked to its entry in a Chinese-English dictionary, and to automatically-generated review exercises. The app estimates the reading proficiency of the user based on a "to-learn" list of vocabulary items. It automatically constructs and maintains this list by tracking the user's dictionary lookup behavior and performance in review exercises. When a user searches for articles to read, search results are filtered such that the proportion of unknown words does not exceed a user-specified threshold.

1 Introduction

"Free voluntary reading" — i.e., recreational reading, or reading for pleasure — promotes reading competence and vocabulary development (Krashen, 2005). Since it plays such an important role in second language acquisition, students benefit from reading a wide range of texts, inside and outside the classroom.

We present a mobile app that facilitates reading among learners of Chinese as a foreign language. The app includes a text database that offers over 500K articles from Chinese Wikipedia, covering a wide range of topics. These articles have been word-segmented. The app provides a supportive reading environment by linking each word to its entry in a Chinese-English dictionary. It also automatically generates review exercises for each word. Further, the app estimates the reading proficiency of the user based on his "to-learn" list of vocabulary items, and maintains this list by tracking user behavior in dictionary lookup and performance in review exercises. When a user searches for articles to read, search results are filtered such that the proportion of unknown words does not exceed a user-specified threshold.

The rest of the paper is organized as follows. Section 2 summarizes the features of the app; these features rely on a user proficiency model, which is presented in Section 3. Section 4 then describes implementation details and evaluates the quality of the review exercises. Section 5 compares this app with other computer-assisted language learning systems. Finally, Section 6 concludes.

2 System Features

The app toggles among three modes — Search, Read, and Review. In addition, there is a "Settings" page for user customization.

2.1 Search Mode

The start page of the app presents a search interface for reading materials (Figure 1a). The user can enter keywords to search for articles on the desired topic. Below the search field, the page displays words that are currently in the user's personal "to-learn" list. By highlighting these keywords, the app steers the user to articles that can reinforce or expand his vocabulary knowledge.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 11–15, Osaka, Japan, December 11-17 2016.

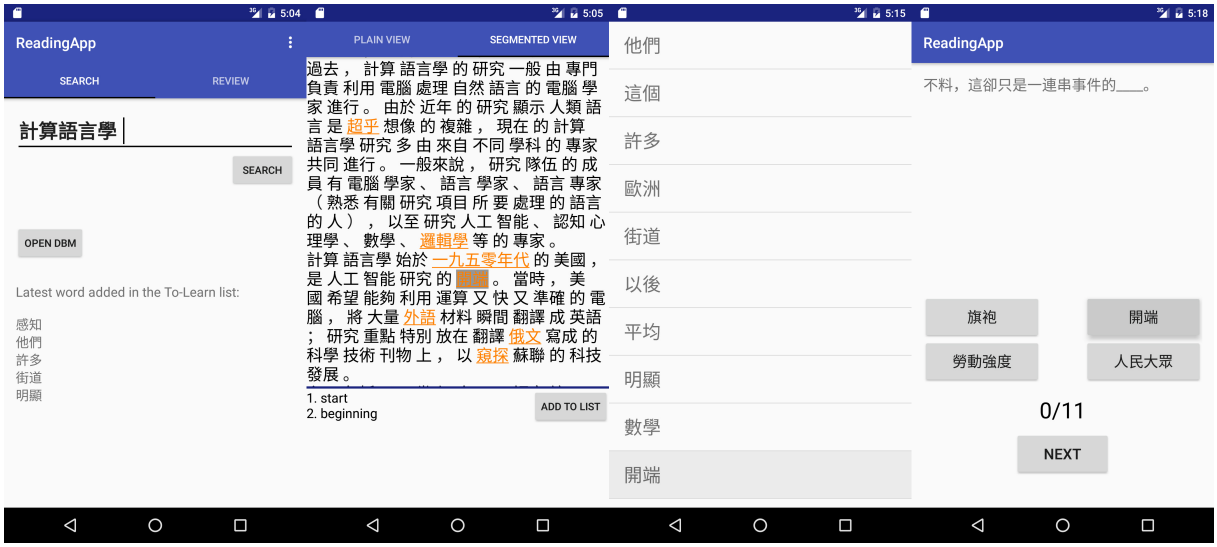

Figure 1: From left to right: (a) **Search Mode** (Section 2.1): The user enters the keyword *ji suan yu yan xue* 'computational linguistics' to search for articles on that topic; below the search field, the user's personal "to-learn" list is shown to provide suggested keywords. (b) **Read Mode** (Section 2.2): One of the retrieved articles is displayed with word segmentation. Words predicted to be unknown to the user are highlighted in orange. The user taps on the word *kai duan* 'beginning' to consult its dictionary entry in English. (c) **Settings page** (Section 2.4): The user views his "to-learn" list, to which the word *kai duan* has been added. (d) **Review Mode** (Section 2.3): A fill-in-the-blank item is offered for the word *kai duan*, with three other distractors.

Previous research suggests that learners need to know 95% to 98% of the words in a text in order to understand it well (Laufer, 1989; Hu and Nation, 2000; Schmitt et al., 2011). While the user should not be overwhelmed with difficulty words, he might nonetheless desire articles that stretch his vocabulary. By default, the search results are filtered such that the proportion of unknown words does not exceed 20%. This percentage can be adjusted by the user to suit his preference (Section 2.4). The app dynamically estimates the user's vocabulary level (Section 3), so that search results keep pace with his increasing proficiency.

2.2 Read Mode

When the user taps on a search result, the app enters the Read Mode and initially displays plain text with no reading aid. The user may choose to request word segmentation and English translations for Chinese words. As shown in Figure 1b, word boundaries are indicated with space, and words predicted to be unknown to the user are highlighted in yellow. When the user taps on a Chinese word, the app shows its English translation at the bottom, and also prompts the user to add the Chinese word to his "to-learn" list.

2.3 Review Mode

At any time, the user can request review exercises for words in his "to-learn" list. We will refer to the word being reviewed as the "target word". The app offers two kinds of exercises:

- ***Translation exercises***: The user is shown the target word and three possible choices of its English translation. One of these choices is the definition extracted from the English dictionary (Section 4.1). The other three are distractors, drawn randomly among other entries in the dictionary. These exercises start with easy words in the "to-learn" list and proceed to the more difficult ones, as estimated by word frequencies in Chinese Wikipedia.

- ***Fill-in-the-blank exercises***: The system randomly draws a sentence from the text database (Section 4.1) that contains the target word. It blanks out the target word and offers four choices. One

choice is the target word itself; the other three are distractors, chosen such that they have the same part-of-speech and similar word frequencies as the target word (Coniam, 1997). Figure 1d shows a fill-in-the-blank item for the word *kai duan* 'beginning'.

If the user picks the right answer for either exercise, the target word is removed from the "to-learn" list, and the user proficiency model is updated (Section 3).

2.4 Settings

On the Settings page, the user can view and adjust three parameters:

- ***"To-learn" list***. The user can view and optionally remove words from the "to-learn" list (Figure 1c).
- ***Vocabulary coverage percentage***: This parameter specifies the minimum percentage of words that must be known to the user in an article, to filter out reading material that would require excessive dictionary lookup. It is set at 80% by default and can be adjusted by the user.
- ***User proficiency level***: The app estimates the user's vocabulary proficiency level (Section 3). The level is shown to the user on a 20-level scale. Level 1 assumes knowledge of the 1000 most frequent words in Chinese Wikipedia. Each subsequent level adds the 1000 next most frequent words, up to Level 20. At this highest level, with the default vocabulary coverage percentage of 80%, the user would be able to read 82.2% of the articles in Chinese Wikipedia. The user can manually adjust his proficiency level in order to obtain easier or more difficult reading material.

3 User Proficiency Model

In order to retrieve texts that challenge the user, yet not overwhelmingly difficult, the app attempts to estimate the user's proficiency level. Automatic proficiency assessment is a difficult task and needs to consider a wide range of factors. Since previous research has shown significant correlation between proficiency and vocabulary level (Laufer and Nation, 1995; Coniam, 1999), we focus on the user's vocabulary size. Specifically, the app estimates the number of Chinese words that the user knows. We rank the words in the user's "to-learn" list according to their frequency in Chinese Wikipedia. The user is then estimated to know all words that have higher frequency than the median word in the list.

A new user is estimated to know 5000 words, the breadth required for the highest level of the *Hanyu Shuiping Kaoshi* (HSK), a widely adopted proficiency test for Chinese. The estimate then dynamically changes according to the set of words in the "to-learn" list. In Read Mode, when the user looks up the English translation of a Chinese word, that word is added to the list. In Review Mode, when the user successfully completes an exercise on a target word, that word is removed from the list. The user can also directly edit the "to-learn" list and/or the proficiency level in the Settings page (Section 2.4).

4 Implementation

4.1 Text Database

We used Solr, a high performance search server that supports full-text search, to construct our database. We extracted a total of 524,543 articles from Chinese Wikipedia to be included in the database. On average, each article has 370 characters and 12 sentences. We segmented all texts with the Stanford Chinese segmenter (Manning et al., 2014). CC-CEDICT, a Chinese to English dictionary with 114,291 entries, supplies English translations for Chinese words in the texts.

4.2 Fill-in-the-blank Exercises

For each target word, the app generates fill-in-the-blank items (Section 2.3). Each item consists of a carrier sentence with a blank, and four choices for the blank. The generation process is as follows:

- ***Carrier sentence selection***: The sentence must contain the target word, and must be between 10 and 20 words long. Further, other words in the sentence must not be more difficult (i.e., have lower frequency) than the target word. Within these constraints, for each target word, ten sentences are selected as carrier sentences.

- ***Distractor generation***: While the key for the item is the target word, the three distractors must be generated. We follow largely the same criteria as Coniam (1997), requiring the distractors to have similar word frequency and the same part-of-speech as the key. It is crucial that a distractor be an unacceptable answer. We evaluated 100 fill-in-the-blank items, randomly chosen from all 20 levels. Overall, 92% of these items had a unique answer (i.e., the target word), while the remaining 8% contained two correct answers.

5 Previous Work

Current systems that support reading in a foreign language mostly focus on English. The user can search for web pages with the *Read-X* tool, which classifies them in real time according to theme and to difficulty level (Miltsakaki and Troutt, 2008); the text is then displayed in the *Toreador* tool, which underlines unknown vocabulary according to the user-specified grade level. Another system, *REAP*, allows the user to search a database of downloaded web pages (Heilman et al., 2008). Similar to our system, *REAP* offers fill-in-the-blank exercises, but they are human-crafted rather than automatically generated.

Fewer systems are available to learners of Chinese as a foreign language. Many focus mainly on teaching characters and words (Shei and Hsieh, 2012). Others, such as *Clavis Sinica* (clavisinica.com) and *Du Chinese* (duchinese.net), use pre-selected texts, vocabulary exercises and translations. The *Smart Chinese Reader* (nlptool.com) allows the user to input any text, and then automatically performs word segmentation and links the words to CC-CEDICT. In addition, it supports automatic sentence translation, and helps the user maintain a "to-learn" word list. Distinct to the systems cited above, our system automatically generates vocabulary review exercises (Section 2.3), and dynamically estimates the user's proficiency level to personalize search results (Section 3).

6 Conclusions and Future work

We have presented an app that offers a reading environment for learners of Chinese as a foreign language. It helps the user search for reading material at an appropriate vocabulary level, and automatically generates review exercises. In future work, we would like to further develop this app in a number of areas. First, we intend to implement more sophisticated criteria for choosing sentences for the review exercises (Kilgarriff et al., 2008). Second, we aim to refine the estimation procedure for the user's vocabulary level (Miltsakaki and Troutt, 2008; Ehara et al., 2012; Ehara et al., 2013). Lastly, we plan to take into account the syntactic complexity of a text when assessing its difficulty level (Heilman et al., 2007).

Acknowledgements

This work is funded by the Language Fund under Research and Development Projects 2015-2016 of the Standing Committee on Language Education and Research (SCOLAR), Hong Kong SAR.

References

David Coniam. 1997. A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests. *CALICO Journal*, 14(2-4):15–33.

David Coniam. 1999. Second Language Proficiency and Word Frequency in English. *Asian Journal of English Language Teaching*, 9:59–74.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining Words in the Minds of Second Language Learners: Learner-specific Word Difficulty. In *Proc. COLING*.

Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. 2013. Personalized Reading Support for Second-Language Web Documents. *ACM Transactions on Intelligent Systems and Technology*, 4(2):31.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proc. NAACL-HLT*.

Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. In *Proc. Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Marcella Hu and I. S. P. Nation. 2000. Unknown Vocabulary Density and Reading Comprehension. *Reading in a Foreign Language*, 13(1):403–430.

Adam Kilgarriff, Mils Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proc. EURALEX*.

Stephen Krashen. 2005. Free Voluntary reading: New Research, Applications, and Controversies. In Gloria Poedjosoedarmo, editor, *Innovative Approaches to Reading and Writing Instruction, Anthology Series 46*, pages 1–9, Singapore. Southeast Asian Ministers of Education Organization (SEAMEO) Regional Language Centre (RELC).

Batia Laufer and P. Nation. 1995. Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, 16(3):307–322.

Batia Laufer. 1989. What Percentage of Text-Lexis is Essential for Comprehension? In Christer Laurén and Marianne Nordman, editors, *Special Language; from Humans Thinking to Thinking Machines*, pages 316–323, Clevedon. Multilingual Matters.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. ACL System Demonstrations*, pages 55–60.

Eleni Miltsakaki and Audrey Troutt. 2008. Real time web text classification and analysis of reading difficulty. In *Proc. Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Norbert Schmitt, Xiangying Jiang, and William Grabe. 2011. The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, 95(i):26–43.

Chris Shei and Hsun-Ping Hsieh. 2012. Linkit: a CALL system for learning Chinese characters, words and phrases. *Computer Assisted Language Learning*, 25(4):319–338.