Personalized Text Retrieval for Learners of Chinese as a Foreign Language

Chak Yan Yeung and **John Lee**
Department of Linguistics and Translation
City University of Hong Kong
chak.yeung@my.cityu.edu.hk, jsylee@cityu.edu.hk

Abstract

This paper describes a personalized text retrieval algorithm that helps language learners select the most suitable reading material in terms of vocabulary complexity. The user first rates their knowledge of a small set of words, chosen by a graph-based active learning model. The system trains a complex word identification model on this set, and then applies the model to find texts that contain the desired proportion of new, challenging, and familiar vocabulary. In an evaluation on learners of Chinese as a foreign language, we show that this algorithm is effective in identifying simpler texts for low-proficiency learners, and more challenging ones for high-proficiency learners.

1 Introduction

Since extensive reading is beneficial for learning a foreign language, students are encouraged to read a wide range of texts, beyond their textbooks and graded readers. Reading materials for language learning need to be carefully selected. On the one hand, if the text is too difficult, it would hinder comprehension and leave the student overwhelmed and discouraged. On the other hand, if the text is too easy, it would not serve to stretch the student's language skills. The best material should be challenging yet highly readable, containing just the right proportion of new, challenging, and familiar vocabulary. Graded readers created with specific language proficiency levels in mind alleviate this problem to some extent, but they provide only a limited amount of materials, and the grade levels are still coarse-grained. It often falls to language teachers to identify appropriate material, but they cannot do so for every individual student.

This paper describes a personalized text retrieval algorithm that helps language learners select the most suitable reading material in terms of vocabulary complexity. The user first rates their knowledge of a small set of words, chosen by a graph-based active learning model. The system trains a complex word identification model (CWI) on this set. Most current CWI approaches assume vocabulary knowledge to be a dichotomy, labeling each word either as "non-complex", if the user is familiar with the word, or as "complex", if the user is not familiar with it (Paetzold and Specia, 2016; Ehara et al., 2014). There are, however, many dimensions of vocabulary knowledge (Richards, 1976). As our first contribution, we take one step in this direction by attempting to identify not only known and unknown words, but also challenging words.

Further, we apply CWI to the text retrieval task, and measure its effect on the vocabulary complexity of retrieved texts from the perspective of a language learner. Specifically, the algorithm attempts to find texts with a minimum proportion of non-complex words and a maximum number of challenging words for the learner. As our second contribution, we show that the algorithm is effective in retrieving simpler texts for low-proficiency learners, and more difficult ones for high-proficiency learners.

The rest of this paper is organized as follows. The next section summarizes previous research in automatic CWI and text retrieval for language learners. Section 3 outlines our approach. Section 4 describes our dataset. Section 5 presents our experimental results. Section 6 concludes the paper.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3448–3455
Santa Fe, New Mexico, USA, August 20-26, 2018.

2 Previous work

2.1 Complex word identification

The complex word identification (CWI) task aims to tag a word as "complex" if it is unfamiliar for a user, and "non-complex" otherwise. In the lexical simplification task, CWI is commonly used for identifying words that should be simplified to improve a user's understanding of a text while avoiding unnecessary simplification (Shardlow, 2013). CWI is also an important part in systems that aim to provide suitable reading material for users. Elhadad (2006), for example, built a system to predict terms in medical literature that were unlikely to be understood by layman readers and provide definitions. This work uses CWI to find texts suitable for learners of Chinese as a foreign language by estimating the proportion of complex, challenging, and non-complex words in a text.

Most previous research has focused on CWI for English as a foreign language. One approach for predicting vocabulary knowledge is by "word sampling" (Laufer and Nation, 1999). Using a ten-level proficiency scale, with 1000 words at each level, this approach samples a fixed number of words from each level and estimates the size of the learner's vocabulary from his/her knowledge of the sampled words.

Word frequencies were found to be the most reliable predictor of word complexity in the 2016 SemEval shared task for CWI in English (Paetzold and Specia, 2016). The best team, which combined various lexicon-based, threshold-based and machine learning voter subsystems, achieved a precision of 0.147 and recall of 0.769. This approach has yet to be evaluated on language learners at different levels of proficiency, since the entire test set was annotated by one learner.

Personalized models for CWI adjust their predictions based on user characteristics. Zeng et al. (2005) showed that demographic features can help improve CWI performance for individual users in the medical domain. Ehara et al. (2012; 2014) performed CWI for individual learners with a two-step approach. In the first step, an active learning approach selects the k most informative nodes, or words, to be annotated by the user (see Section 3.1 for details). The user rates their knowledge of these k words on a five-point scale (Table 1). In the second step, the system uses Local and Global Consistency (Zhou et al., 2004), a label propagation algorithm, to train an independent classifier for each user. The algorithm performs binary classification on the nodes to predict whether a user knows a word or not. The assumption behind this algorithm is that two nodes connected by a heavily weighted edge should have similar labels, and more heavily weighted edges should propagate more labels. The best model, where k=50, achieved 76.44% accuracy on a dataset of Japanese learners of English.

Score	Description	Label (Ehara)	Label (this study)
1	Never seen the word before	complex	complex
2	Probably seen the word before	complex	complex
3	Absolutely seen the word before but do not know its meaning, or tried to learn the word before but forgot its meaning	complex	complex
4	Probably know, or able to guess, the word's meaning	complex	challenging
5	Absolutely know the word's meaning	non-complex	non-complex

Table 1: Five-point scale for rating vocabulary knowledge, and the mapping to CWI labels used by Ehara et al. (2012; 2014) and in this study.

Recently, CWI for other languages has begun to receive more attention. The latest CWI shared task, for example, featured multilingual datasets, covering English, German, Spanish, and French (Yimam et al., 2018). To the best of our knowledge, there has so far been only one reported study on CWI for Chinese as a foreign language (CFL) (Lee and Yeung, 2018). They followed Ehara et al. (2012; 2014) in selecting 50 most informative words to be rated by CFL learners. They then used a support vector machine (SVM) classifier to train a CWI classifier for each individual learner based on his/her 50-word training set. The classifier used features based on word difficulty levels in two assessment scales for CFL, namely the Test of Chinese as a Foreign Language (TOCFL) (Zeng, 2014) and the Hanyu Shuiping Kaoshi (HSK) (Hanban, 2014). Details on the features are provided in Section 3.2. On a dataset with seven CFL learners, this approach achieved 78% CWI accuracy, outperforming the label propagation method. In this study, we adopt this classification approach and extend it to three-way, so that we may use not just the proportion of complex and non-complex words, but also challenging words, as criteria for text retrieval.

2.2 Text retrieval for language learners

Past research has shown that the optimal reading material should fit the learner in terms of vocabulary and grammatical complexity, as well as personal interests (Heilman et al., 2007). In this study, we focus on the proportion of non-complex words in a text.

Several studies have been conducted to examine the proportion of words in a text that a learner needs to know to facilitate reading comprehension. There is no consensus on the "optimal" proportion of known words in a text. Laufer and Nation (1999), Liu and Nation (1985) and Hirsh et al. (1992) showed that a student could read and understand a text with ease if 95% to 98% of the words were known. If the goal is to stretch the student's vocabulary, the threshold can be made lower. Indeed, in the first 5 books of the *New Practical Chinese Reader*, a popular textbook series for Chinese as a foreign language (CFL), the proportion of difficult words ranges from 9% to 31% (Liang and Song, 2009).

3 Approach

We propose a text retrieval algorithm that aims to help language learners select the most suitable reading material in terms of vocabulary complexity. The algorithm has three components. First, it employs a graph-based active learning model to select a small set of words as training set (Section 3.1). After the user has rated his/her knowledge of these words, the system trains a complex word identification model on this set (Section 3.2). It then applies the model to find texts that contain the desired proportion of complex, challenging, and non-complex words (Section 3.3). We now provide details on each of these components.

3.1 Training set creation

The system performs graph-based active learning to select a small number of words to be included in the training set. The entire vocabulary is first organized as a multiple complete graph. Nodes correspond to words, and edge weights reflect the similarity between the frequency ranks of the words. The assumption is that words with similar frequency ranks are known to learners whose familiar words are similar to each other. Using Error Bound Minimization (Gu and Han, 2012), a non-interactive graph-based active learning algorithm, the system selects the k most informative nodes from the vocabulary graph in a non-interactive way. This algorithm selects nodes that are globally important, based on the number of edges. Further, the nodes must not be heavily connected to previously sampled nodes. Following previous work (Lee and Yeung, 2018; Ehara et al., 2014), we set k as 50. While a larger k can yield better performance, burdening users with a large amount of vocabulary annotation would be undesirable in practice.

3.2 Complex word identification

The CFL learners scored their knowledge on the 50 words in the training set with the five-point scale in Table 1. Following Ehara et al. (2014) and Lee and Yeung (2018), we asked the learners to score their knowledge of words in isolation, rather than in context such as in the CWI shared task.

The context of a word provides clues for learners to guess its meaning, and thus affects how they score their knowledge of a word. Even if the learner is able to guess a word in one context, the same is not guaranteed in another context since the content and the density of new words in each text is different. Since the CWI model in our system is intended for text selection, we did not wish to assume any one particular context when determining the learner's knowledge of a word. To more accurately judge the learners' ability to understand different reading materials, we asked them to consider each word in isolation.

For each individual user, we trained a classifier based on his/her annotation on the 50-word training set to perform CWI on all Chinese words. In a departure from previous CWI studies, we attempted three-way classification, labelling each word as "non-complex", "challenging" or "complex" (see Table 1). We experimented with a feature set similar to that of Lee and Yeung (2018):

- **WordList**: The difficulty level of the word (1, 2, 3, 4, 5, or 6), according to the HSK and TOCFL guidelines. HSK was used as the basis for this feature. For words not covered by HSK, we mapped their TOCFL level to HSK.
- **HSK-char-max**: A Chinese word may contain multiple characters. This feature takes the maximum level among the characters in the word (1, 2, 3, 4, 5, or 6), according to the HSK guidelines.
- **Freq**: The frequency of the word in Chinese Wikipedia.
- **LearnerFreq**: The frequency of the word in the *Jinan Corpus of Learner Chinese* (JCLC) (Wang et al., 2015).
- **LearnerFreq-char-min**: For each character in the word, we compute its frequency count in JCLC. This feature takes the minimum frequency.

3.3 Text retrieval

Our system allows the user to specify the minimum (estimated) proportion of words in the retrieved text that he/she should be able to understand. This parameter can be set to an arbitrary percentage between 0 and 100, but in consideration of the manual effort needed for evaluation, we set this parameter to 80% in our experiment.

Ehara et al. (2012) assumed that learners could only understand the words that received a score of 5 on the five-point scale in Table 1, and that they did not know any word with a score of 4 or below. The aim of our system, however, is to stretch the student's language skills. Taking lexical guessing in context into account, we consider both non-complex and challenging words, scored five and four on the five-point scale respectively, as words that the user should be able to understand.

To stretch the learner's language skills while maintaining a high level of comprehension, the system retrieves all documents that satisfy the minimum threshold of 80% non-complex words, and then ranks them from the highest proportion of challenging words to the lowest.

4 Datasets

4.1 Complex word identification

We used the same training set of 50 words and test set of 550 words as Lee and Yeung (2018). We retained the original annotations of the seven CFL learners, and included the annotations of an additional learner. The learners labelled each word in the datasets on a five-point scale (see Table 1). We considered a word to be "non-complex" if it was scored five, "challenging" if it was scored four, and "complex" otherwise.

The numbers of "non-complex", "challenging", and "complex" words of the test set and training set for each learner are shown in Table 2. We divided the eight learners into two groups. The four who knew less than 150 of the 550 words were designated as "Low-Proficiency", and the other four as "High-Proficiency".

4.2 Text retrieval

We compiled a collection of texts from Chinese Wikipedia, Chinese Internet Corpus (Sharoff, 2006), and *Duan Mei Wen*[1]. For each text, we extracted the first n sentences such that the total character count was between 200 to 225. A total of 38,881 texts were included as candidates for the text retrieval algorithms.

We evaluated three text retrieval methods. Our baseline method randomly selects texts. The other two methods use the two CWI models that achieved the best performance on low-proficiency learners and

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

User	Training Set			Test Set		
	non-complex	challenging	complex	non-complex	challenging	complex
1	11	2	37	68	25	457
2	7	9	34	78	91	381
3	22	10	18	296	38	216
4	14	8	28	113	114	323
5	17	8	25	217	116	217
6	13	15	22	188	109	253
7	14	8	28	146	75	329
8	20	13	17	268	135	147

Table 2: The numbers of non-complex, challenging, and complex words in the training sets and test sets.

high-proficiency learners (Table 3), respectively, coupled with the retrieval criteria described in Section 3.3. We retrieved the top five texts for each of these three methods.

We asked four of the eight learners — two low-proficiency learners and two high-proficiency learners — to annotate the fifteen retrieved texts with the same five-point scale used in the CWI experiment. They were asked to highlight the complex words in red and the challenging words in blue.

5 Experimental results

5.1 Complex word identification

Following Ehara et al. (2014) and Lee and Yeung (2018), we trained SVM classifiers and logistic regression (LR) classifiers. We used the implementation in scikit-learn (Pedregosa et al., 2011), and we tried all combinations of the features listed in Section 3.1. The results are shown in Table 3.

Features	Model	All	Low-proficiency	High-proficiency
WordList	SVM	64.18%	**70.96%**	57.41%
WordList	LR	63.66%	70.41%	56.91%
WordList + HSK-char-max	SVM	63.57%	69.68%	57.46%
WordList + HSK-char-max	LR	58.82%	65.46%	52.18%
Freq	SVM	40.52%	36.82%	44.23%
Freq	LR	56.23%	62.82%	49.64%
LearnerFreq	SVM	57.02%	64.50%	49.55%
LearnerFreq	LR	62.93%	70.09%	55.78%
LearnerFreq + Freq	SVM	51.11%	60.64%	41.59%
LearnerFreq + Freq	LR	60.64%	66.73%	54.55%
LearnerFreq + LearnerFreq-char-min	SVM	43.14%	50.46%	35.82%
LearnerFreq + LearnerFreq-char-min	LR	64.27%	69.87%	58.68%
LearnerFreq + LearnerFreq-char-min + WordList	SVM	54.61%	56.36%	52.86%
LearnerFreq + LearnerFreq-char-min + WordList	LR	**64.68%**	70.59%	**58.77%**

Table 3: The average accuracies of 3-way complex word identification for all eight learners, low-proficiency learners, and high-proficiency learners.

The LR model with LearnerFreq, LearnerFreq-char-min, and WordList features achieved the highest accuracy, at 64.68%. It slightly outperformed the LR model with LearnerFreq and LearnerFreq-char-min features, at 64.27%, and the SVM model with WordList features, at 64.18%. In general, CWI performance is better on low-proficiency learners than on high-proficiency learners.

The WordList features seem particularly useful for capturing the limited vocabulary of beginners, with the word lists covering most of the words that they know. This hypothesis is consistent with the fact that the SVM model with WordList features achieved the best performance for the low-proficiency learners, at 70.96%. For high-proficiency learners, however, the word lists are insufficient in modelling their larger vocabularies, and have to be augmented with frequency features in order to cover a wider range of words. The LR model with LearnerFreq, LearnerFreq-char-min, and WordList features achieved the highest accuracy for high-proficiency learners, at 58.77%.

5.2 Text retrieval

We evaluated three text retrieval methods:

- **WordList**: The SVM model with WordList features, which achieved the best CWI results for low-proficiency learners.
- **Learner + WordList**: The LR model with LearnerFreq, LearnerFreq-char-min, and WordList features, which achieved the best CWI results for high-proficiency learners,
- **Baseline**: Random selection.

Experimental results for high- and low-proficiency learners are shown in Tables 4 and 5, respectively. To prevent outliers from skewing the results, we removed the texts with the highest and the lowest proportion of non-complex words for each user.

Model	**Non-complex words**	**Challenging words**
WordList	**90.18%** (estimated: 85.06%)	7.84% (estimated: 18.13%)
Learner + WordList	**87.87%** (estimated: 84.62%)	6.13% (estimated: 23.39%)
Random selection	95.36%	4.82%

Table 4: The average proportion of non-complex and challenging words in text retrieved for high-proficiency learners.

Model	**Non-complex words**	**Challenging words**
WordList	**80.60%** (estimated: 82.66%)	5.91% (estimated: 38.24%)
Learner + WordList	**86.26%** (estimated: 83.91%)	3.76% (estimated: 20.45%)
Random selection	72.91%	6.09%

Table 5: The average proportion of non-complex and challenging words in text retrieved for low-proficiency learners.

For high-proficiency learners, randomly selected texts were relatively easy to read. This can be seen in the high proportion of non-complex words (95.36%) in texts returned by the baseline method. In comparison, the WordList model and Learner+WordList model both retrieved more difficult texts, with 90.18% and 87.87% of non-complex words, respectively. Thus, these models were able to find texts that presented more new vocabulary to advanced learners, while still exceeding the minimum comprehension rate of 80

For low-proficiency learners, experimental results showed the opposite effect. The baseline method retrieved rather difficult texts, in which only 72.91% of the words were non-complex, below the desired minimum rate of 80%. In contrast, the WordList model and Learner+WordList model retrieved texts with 80.60% and 86.26% of non-complex words, respectively. Thus, they succeeded in finding more readable texts, with above 80% comprehension rate, for the beginners.

For both the WordList model and the Learner+WordList model, the proportion of challenging words in the retrieved texts was lower than estimated. As shown in Table 6, among the words estimated to be challenging by the two best CWI models, most (63.11% and 75.15%, respectively) were scored as non-complex by the learners. The discrepancy may be attributable to the fact that learners, with the benefit of context while reading a text, can better guess the meaning of a word that would otherwise be challenging. This phenomenon can be even more clearly seen by comparing the scores on the same words by the same learner in the CWI dataset (Section 4.1) and the text retrieval dataset (Section 4.2). Indeed, among the words that were annotated as challenging in isolation in the CWI dataset, 68% were annotated as non-complex in the text retrieval dataset by the same learners. Similarly, among the words that were annotated as complex in the CWI dataset, 36.67% were annotated as non-complex in the text retrieval dataset.

We conducted further analysis on the kinds of words that learners are more likely to find non-complex when reading a text. We found that "challenging" or "complex" single-character words (e.g. 以 *yǐ* (with), 称 *chēng* (named)) were more likely to be considered "non-complex" in context. Words related to the theme of the texts (e.g. 口音 *kǒu yīn* (accent) in a text about dialects and 政治 *zhèng zhì* (politic) in a text about democracy) were also more likely to be changed from "challenging" or "complex" to "non-complex".

Gold label	**All**	**Low-proficiency**	**High-proficiency**
WordList			
Complex	26.99%	32.16%	21.83%
Challenging	9.90%	6.04%	13.76%
Non-complex	63.11%	61.79%	64.42%
Learner + WordList			
Complex	17.44%	14.83%	20.06%
Challenging	7.41%	2.51%	12.30%
Non-complex	75.15%	82.67%	67.64%

Table 6: Gold label for words estimated to be challenging in the text retrieval dataset (Section 4.2).

6 Conclusion

Reading materials for language learning need to be carefully selected. The best material should be challenging yet highly readable, containing just the right proportion of new and challenging vocabulary. This paper describes a personalized text retrieval algorithm that aims to help language learners select the most suitable reading material in terms of vocabulary complexity.

The system first employs a graph-based active learning model to select a small set of informative words. Each user rates his/her knowledge of these words to create a training set for complex word identification (CWI). The system then trains a personalized CWI model, and applies it to find texts that contain the desired proportion of complex, challenging, and non-complex words. We evaluated this algorithm on learners of Chinese as a foreign language, and showed its effectiveness in identifying easier texts for low-proficiency learners, and more challenging ones for high-proficiency learners.

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. We thank Dr. Lis Pereira for assisting with the experiments.

References

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: learner-specific word difficulty. *Proceedings of COLING 2012*, pages 799–814.

Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. 2014. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1384.

Noemie Elhadad. 2006. Comprehending technical texts: Predicting and defining unfamiliar terms. In *AMIA annual symposium proceedings*, volume 2006, page 239. American Medical Informatics Association.

Quanquan Gu and Jiawei Han. 2012. Towards active learning on graphs: An error bound minimization approach. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 882–887. IEEE.

Hanban. 2014. *International curriculum for Chinese language education*. Beijing Language and Culture University Press, Beijing, China.

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 *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467.

David Hirsh, Paul Nation, et al. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a foreign language*, 8:689–689.

Batia Laufer and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language testing*, 16(1):33–51.

John Lee and Chak Yan Yeung. 2018. Automatic prediction of vocabulary knowledge for learners of chinese as a foreign language. In *Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*.

Ji-hua Liang and Shao-li Song. 2009. Construction of an approach for counting chinese graded words and characters——a tool for assessing difficulty level of word in chinese language teaching materials writing system. *Modern Educational Technology*, 7:024.

Na Liu and ISP Nation. 1985. Factors affecting guessing vocabulary in context. *RELC journal*, 16(1):33–42.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jack C Richards. 1976. The role of vocabulary teaching. *TESOL quarterly*, pages 77–89.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98.

Maolin Wang, Shervin Malmasi, and Mingxuan Huang. 2015. The jinan chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *International Symposium on Biological and Medical Data Analysis*, pages 184–192. Springer.

W. Zeng. 2014. Huayu baqianci ciliang fenji yanjiu (classification on chinese 8000 vocabulary). *Huayu Xuekan*, pages 6: 22–33.

Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328.