Automatic Prediction of Vocabulary Knowledge for Learners of Chinese as a Foreign Language

John Lee
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong
jsylee@cityu.edu.hk

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

***Abstract*— Since extensive reading is beneficial for learning a foreign language, students are encouraged to seek additional reading materials from sources beyond their textbooks. The materials should be difficult enough to stretch the student's language proficiency, but not too difficult as to hinder comprehension. A complex word identification (CWI) system can identify texts that optimize these criteria by estimating the student's proficiency level. We present a personalized CWI model for Chinese as a foreign language. This model predicts whether the student knows a Chinese word or not, based on a small training set from the student. In empirical evaluation, a support vector machine (SVM) classifier with features based on graded vocabulary lists yielded the best performance, outperforming a label propagation approach that is state-of-the-art for personalized CWI for English.**

Keywords— complex word identification, computer-assisted language learning, Chinese as a foreign language, vocabulary modeling

I. INTRODUCTION

"Free voluntary reading" — i.e., recreational reading, or reading for pleasure — serves as a major source of reading competence and vocabulary development [6]. Since reading plays such an important role in second language acquisition, students benefit from reading 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 unfamiliar, or complex, words.

Our goal is to develop a system that helps language learners search for reading materials with suitably difficult vocabulary. Such a system needs to be able to estimate the vocabulary knowledge of individual users, in order to find texts that have the desired percentage of complex words. This paper evaluates a simple but effective approach for this task in the context of learning Chinese as a foreign language (CFL). We use a support vector machine (SVM) classifier to perform personalized complex word identification (CWI) for individual CFL learners, based on a small training set annotated by each learner. Empirical evaluations show that it outperforms a label propagation method [2], which yields state-of-the-art performance for personalized CWI for English as a foreign language.

The rest of this paper is organized as follows. The next section summarizes previous research in automatic CWI. Section 3 outlines our approach for estimating vocabulary knowledge. Section 4 describes our dataset. Section 5 presents our experimental results. Section 6 concludes and discusses future work.

II. COMPLEX WORD IDENTIFICATION

The task of complex word identification (CWI) is to label each word as "non-complex", if the user is familiar with the word, and as "complex", if the user is not familiar with it. CWI is useful for computer-assisted language learning because it can help select suitable reading materials for students of foreign languages. For example, it can restrict search results to documents where the learner knows at least 95% to 98% of the words in a text, to ensure the student can understand the material with ease [5]. 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% [9]. To the best of our knowledge, no study on CWI for CFL has been reported.

One approach for predicting vocabulary knowledge is by "word sampling" [7]. Using a ten-level proficiency scale, with 1000 words at each level, this approach samples a fixed number of words as the training set, and then labels unseen words based on their proximity to these words.

Word frequencies were found to be the most reliable predictor of word complexity in the 2016 SemEval shared task for CWI in English [10]. 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.

Zeng et al. showed that demographic features can help improve CWI performance for individual users in the medical domain [13]. Ehara et al. performed CWI for individual learners with an active learning approach followed by label propagation [1][2]. The best model achieved 76.44% accuracy on a dataset of Japanese learners of English. In this approach, 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

978-1-5386-4543-7/18/$31.00 ©2018 IEEE

assumption is that words with similar frequency ranks are known to learners whose familiar words are similar to each other. The model then performs the two following steps.

Graph-based Active Learning. First, it uses Error Bound Minimization [3], a non-interactive graph-based active learning algorithm, to select the *k* most informative nodes from the vocabulary graph in a non-interactive way, i.e., without using human labels during the learning process. The values of *k* range from 10 to 50. 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.

Label Propagation. After selecting the nodes by active learning, the system uses Local and Global Consistency [15], 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 initial seed of labeled nodes correspond to the set of selected nodes by active learning, labeled by the user as complex or non-complex. The labels are then propagated to the unlabeled nodes based on edge weights. 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.

III. APPROACH

We first discuss existing resources and guidelines for learners of Chinese as a foreign language (Section 3.1), from which we derive features to train classifiers for complex word identification (Section 3.2). We then describe the baselines and oracle to be used in our evaluation (Section 3.3).

A. Assessment guidelines for Chinese as a foreign language

For Chinese as a foreign language (CFL), the two major assessment scales are the *Test of Chinese as a Foreign Language* (TOCFL) [14] and the *Hanyu Shuiping Kaoshi* (HSK) [4]. Both scales have six levels and can be mapped to the *Common European Framework of Reference for Languages*, a global standard for measuring foreign language proficiency. The HSK guidelines provide a character list and a vocabulary list for each level, covering a total of 9,600 vocabulary items. The TOCFL guidelines also provide similar vocabulary lists, covering a total of 8,000 vocabulary items. We used the HSK as the basis for features that involve the difficulty level of a word (Section 3.2). For words not covered by HSK, we mapped their TOCFL level to HSK.[1]

B. Features

We approach the task of complex word identification (CWI) by training a classifier for each individual user. The classifier accepts any Chinese word as input; its output is "non-complex" if the user knows the word, and "complex" if the user does not. We implemented the following features:

- **HSK+TOCFL**: The difficulty level of the word (1, 2, 3, 4, 5, or 6), according to the HSK and TOCFL guidelines (Section 3.1).
- **HSK-char-max**: A Chinese word may contain multiple characters, potentially at various levels of difficulty. This feature takes the maximum level among the characters in the word (1, 2, 3, 4, 5, or 6), according to the HSK guidelines.
- **HSK-char-min**: Same as above, except that the minimum of the difficulty levels is taken.
- **Freq**: The frequency of the word in a standard Chinese corpus.
- **Freq-char-max**: For each character in the word, we compute its frequency count in the standard Chinese corpus. This feature takes the maximum frequency.
- **Freq-char-min**: Same as above, except that the minimum is taken.
- **LearnerFreq**: The frequency of the word in a learner Chinese corpus.
- **LearnerFreq-char-max**: For each character in the word, we compute its frequency count in the learner Chinese corpus. This feature takes the maximum frequency.
- **LearnerFreq-char-min**: Same as above, except that the minimum is taken.

C. Baselines and oracle

In our experiments, we contrast the classifier approach described above with two baselines. The first is the Majority baseline, which always predicts the label that is assigned to the majority of the words in the training set. The **Majority** baseline can be a very strong baseline for low-proficiency language learners, who have limited vocabulary knowledge. The second baseline is **Label Propagation**, as outlined in Section 2.

Further, to benchmark the difficulty of the task, we implemented the HSK+TOCFL **Oracle**. This oracle considers six "typical" learners, whose vocabulary knowledge conforms exactly to one of the six levels in the HSK and TOCFL guidelines (Section 3.1); specifically, if the learner is at level N, then he or she knows precisely all those words in the corresponding vocabulary lists of HSK and TOCFL from level 1 up to level N, and no other word. The oracle makes predictions according to the "typical" learner that would optimize classification accuracy on the test set.

IV. DATA

To derive word frequency statistics for the Freq feature, we used a corpus of 9.2 million sentences from Chinese Wikipedia. For the LearnerFreq feature, we used the *Jinan Corpus of Learner Chinese* (JCLC) [12], which contains 6 million Chinese characters written by students from over 50 different native language backgrounds. We performed word

[1] http://www.fachverband-chinesisch.de/chinesischindeutschland/pruefungen/index.html

segmentation on both corpora with the Stanford Chinese parser [8].

We constructed our training set and test set from the *Lexical Lists for Chinese Learning in Hong Kong*, compiled by the Hong Kong Education Bureau.[2] This list, consisting of 9,000 Chinese words, defines the expected Chinese vocabulary proficiency for students graduating from secondary schools in Hong Kong. Using the Graph-based Active Learning method described in Section 2, we selected the most informative nodes to generate five training sets, with 10, 20, 30, 40, and 50 words, following Ehara et al. [2]. While we could in principle use larger training sets, it would in practice be undesirable to burden users with a large amount of vocabulary annotation before allowing them to use the system.

As test set, we drew 550 words from the same list that did not overlap with the training set. We randomly selected words that span different levels of difficulty, as measured by their frequency counts in Chinese Wikipedia.

We asked seven subjects, all CFL learners, to label each word in these datasets on a five-point scale: (1) Never seen the word before; (2) Probably seen the word before; (3) Absolutely seen the word before but do not know its meaning, or tried to learn the word before but forgot its meaning; (4) Probably know, or able to guess, the word's meaning; and (5) Absolutely know the word's meaning. Following Ehara et al. [1], we consider a word to be non-complex if it is scored five, and complex otherwise.

V. Experimental Results

Similar to Ehara et al. [2], we trained SVM classifiers and logistic regression (LR) classifiers, using the implementation in scikit-learn [11], with all combinations of the features listed in Section 3.2. The classifier trained with the HSK+TOCFL feature alone yielded the best performance, followed by LearnerFreq and Freq. Table 1 shows their accuracy levels, in comparison to the baselines and oracle (Section 3.3).

With a small training set of just 10 words, the Label Propagation method performed at 70.8% accuracy. Its performance slightly degraded for the training sets with 20 and 30 words. The best performance, 71.4%, was reached for the training sets with 40 and 50 words. As a point of reference, this accuracy rate was about 5% lower than the equivalent result for English vocabulary knowledge prediction [2].

The Majority baseline predicted all words to be unknown. Since our subjects had relatively low proficiency in Chinese, this baseline yielded a strong performance of 72.2% accuracy, slightly outperforming the Label Propagation method (71.4%), as well as the SVM classifier based on word frequency in Chinese Wikipedia (Freq (SVM)), whose accuracy was 70.4%. Using word frequencies in the *Jinan Corpus of Learner Chinese* (JCLC), however, led to more accurate results than Chinese Wikipedia. An LR classifier trained on JCLC (LearnerFreq (LR)) reached 76.3% accuracy, above the baseline. Despite the smaller corpus size, word usage statistics from texts produced by language learners themselves appear to align more closely to their vocabulary knowledge.

The best performance was given by the SVM classifier based on word difficulty levels in the HSK and TOCFL scales (HSK+TOCFL (SVM)), which achieved an accuracy of 78.0%. This result confirms the quality of these scales, which were authored by experts in Chinese language pedagogy. Combining the HSK+TOCFL feature with other features (Section 3.2) resulted in slight degradation in performance.

TABLE I. Accuracy in complex word identification using various methods and training sets.

Method	Training set size	Accuracy
Majority baseline	n/a	0.722
Label propagation	10	0.708
Label propagation	20	0.649
Label propagation	30	0.686
Label propagation	40	0.714
Label propagation	50	0.714
Freq (SVM)	50	0.704
LearnerFreq (LR)	50	0.763
HSK+TOCFL (SVM)	**50**	**0.780**
HSK+TOCFL (Oracle)	n/a	0.791

This may not be unexpected since the accuracy of the HSK+TOCFL approach already lay within 1.1% of the oracle (HSK+TOCFL (Oracle)), suggesting relatively small room for further performance gain with larger training sets.

VI. Conclusions and Future Work

We have reported the first evaluation on the task of complex word identification (CWI) for learners of Chinese as a foreign language (CFL). We used a graph-based active learning algorithm to select words for the training set, and evaluated a classification approach based on features derived from word frequencies in both standard and learner corpora, and from existing CFL assessment scales. We compared this approach with a label propagation method that is state-of-the-art in personalized CWI for English as a foreign language [2]. We obtained the best result with an SVM classifier trained on features based on word difficulty levels in the HSK and TOCFL assessment scales. This classifier achieved 78.0% accuracy in predicting whether a CFL student knows a word, outperforming the label propagation approach.

In future work, we intend to further improve CWI accuracy by investigating linguistic features beyond word frequency counts. We also plan to incorporate the CWI model as a component in a CFL document retrieval system, to enable the system to identify reading materials that are optimal for the vocabulary proficiency level of the student.

Acknowledgments

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),

[2] http://www.edbchinese.hk/lexlist/

Hong Kong SAR. We thank Dr. Lis Pereira for assisting with the experiments.

REFERENCES

[1] Y. Ehara, N. Shimizu, T. Ninomiya, and H. Nakagawa. "Personalized reading support for second-language web documents by collective intelligence," in *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI)*, Hong Kong, China, 2010, pp. 51–60.

[2] Y. Ehara, Y. Miyao, H. Oiwa, I. Sato and H. Nakagawa. "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)*. Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1374–1384.

[3] Q. Gu and J. Han. "Towards active learning on graphs: An error bound minimization approach," in *Proceedings of the IEEE International Conference on Data Mining (ICDM),* 2012.

[4] Hanban. *International Curriculum for Chinese Language and Education. Beijing Language and Culture University Press*, Beijing, China, 2014.

[5] M. H-C. Hu and P. Nation. "Unknown vocabulary density and reading comprehension." *Reading in a Foreign Language*, 13 (1), 2000, pp. 403–430.

[6] S. Krashen. "Free voluntary reading: New research, applications, and controversies." *Innovative Approaches to Reading and Writing Instruction, Anthology Series 46*, SEAMEO Regional Language Centre, Singapore, 2005, pp. 1–9.

[7] B. Laufer and P. Nation. "A vocabulary-size test of controlled productive ability." *Language Testing*, 16(1), 1999, pp. 33-51.

[8] R. Levy and C. D. Manning. "Is it harder to parse Chinese, or the Chinese Treebank?" in *Proc. 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

[9] S. Liang and J. Song. "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 [in Chinese]." *Modern Educational Technology 现代教育*, 19(7), 2009.

[10] G. H. Paetzold and L. Specia. "SemEval 2016 task 11: complex word identification," in *Proc. 10th International Workshop on Semantic Evaluation (SemEval)*, 2016.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg and J. Vanderplas. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*, 12, 2011, pp. 2825-2830.

[12] M. Wang, S. Malmasi and M. Huang. "The Jinan Chinese Learner Corpus," in *Proc. BEA Workshop*, 2015.

[13] Q. Zeng, E. Kim, J. Crowell and T. Tse. "A text corpora-based estimation of the familiarity of health terminology." *ISBMDA 2005*, LNBI 3745, 2005, pp. 184–192.

[14] W. Zeng. "Huayu baqianci ciliang fenji yanjiu (Classification on Chinese 8000 Vocabulary." *Huayu Xuekan 华语学刊*, 6, 2014, pp. 22–33.

[15] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Scholkopf. "Learning with local and global consistency." *Advances in Neural Information Processing Systems*, 2004, pp. 321–328.