Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Automatic Vocabulary Study Map Generation by Semantic Context and Learning Material Analysis

Brendan FLANAGAN[a*], Mei-Rong Alice CHEN[a], Louis LECAILLIEZ[b], Rwitajit MAJUMDAR[a], Gökhan AKÇAPINAR[ac], Patrick OCHEJA[b] & Hiroaki OGATA[c]
[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b]*Graduate School of Informatics, Kyoto University, Japan*
[c]*Department of Computer Education & Instructional Technology, Hacettepe University, Turkey*
*flanagan.brendanjohn.4n@kyoto-u.ac.jp

Abstract: Learning English as a foreign language is a core part of K-12 education for many countries in which English is not the main spoken language, and especially in Asia. One of the fundamental tasks that students encounter is to learn vocabulary that is a part of the assigned curriculum. These are often sourced from reference materials or assigned vocabulary lists and may not consider the learner's current proficiency or the semantic context of words that were recently learnt. By suggesting vocabulary that have similar proficiency or semantic contexts to what a student has recently studied could improve and support vocabulary learning. In this paper, we propose a method for recommending words that have similar difficulty and semantic context with previous words learnt based on the analysis of prescribed textbooks for Japanese junior high school students. This research could be used to guide a student learning English by helping them select a sequence of vocabulary that is appropriate.

Keywords: English as a foreign language, vocabulary learning, semantic context, proficiency

1. Introduction

When studying vocabulary, students are often faced with the challenge of selecting appropriate words to learn at their current level of proficiency. Often students will refer to reference materials or an assigned vocabulary list that are part of their current curriculum. However, these materials do not consider the learners previously learnt vocabulary, which can give insight into the semantic context of their current study.

Nation & Hunston (2013), suggest that learning from context is one of the most important sources when studying vocabulary, and should be an integral part of second language learning. This highlights the strength of the role that context plays in helping learners not only understand the meaning of a word from its surrounding sentence, but also the similarities in contexts of words which are close to each other semantically. Wolter (2006) investigated the differences in native language lexical networks and the learner's perception of the foreign language lexical network. It was found that learners often rely on their knowledge from the native language lexical network. Borodkin et al. (2016) found that foreign language learners do not possess lexical networks that are as well-organized as the network of their native language. Therefore, learning vocabulary from a recommended vocabulary map might help improve the lexical network of the EFL or ESL learners.

In this paper, we propose a method for recommending English vocabulary that a learner could study based on a map of the whole curriculum. The map has been organized with the intention of recommending words that are closely related both semantically and according to the student's current level of acquisition. The analysis of three English as a foreign language textbooks that are widely used at Japanese junior high schools are examined as case studies of the proposed method.

Previous research into the use of semantic context analysis in language learning has focused on the substitution of difficult or unknown words to support reading. Azab et al. (2015) analyzed existing semantic maps, such as: WordNet (Miller, 1995), Roget, and Encarta, to find synonymous that could be used as substitutes to describe words the learner could not understand with reading text with a web browser. While similar methods could be employed to the problem presented in this paper, there are numerous words that excluded from such semantic maps, and the relations are not derived from the actual context in which the word is used. To overcome these limitations, we propose that a word2vec corpus trained using the actual contexts of words would be better suited to the task of recommending vocabulary for learners.

2. Method Overview

2.1 Data Collection: Learning Materials

In this research, we focus on the analysis of three-year levels of learning materials that are used extensively in English Language classes at Japanese Junior High-Schools. A list of prescribed vocabulary is available for each textbook, and was used as a mask to focus the analysis in this paper. However, it should be noted that it is also possible to extract such information automatically from a set of textbooks by comparing the unique words in each book. An overview of the books and vocabulary is shown in Table 1.

Table 1
An overview of the three different levels of EFL learning materials that were targeted in this paper.

Junior High School Year Level	*Vocabulary Studied*	*Pages*
1	*389*	*124*
2	*225*	*116*
3	*341*	*106*

An ordered list of the vocabulary was extracted from each of the textbooks by finding the first mention of a word in the book. If words occurred on the same page, then words that were higher up the page were considered to have occurred before words that were lower down the page as the book is in single column layout.

2.2 Data Collection: Semantic Similarity Corpus and Vector Training

Word2vec was proposed by Mikolov et al. (2013) to effectively learn word embedding representation by examining the context of surrounding words in sample sentences. In essence, the resulting model is trained to predict a word in a sentence based on the surrounding words within a specified window. The weights of the resulting model are then used as a vector representation of the predicted word in a semantic space. Therefore, words that are around a similar area within the vector space are words that usually occur in similar contexts, and therefore have a similar meaning. This technique has been successfully applied to several semantic problems in previous research (Mikolov et al., 2013; Lau & Baldwin, 2016), and could be used to find semantically similar words for language learners to study.

For the purpose of this research, we trained a word2vec vector space model using a corpus based on the full collection of English Wikipedia. The trained model contains a total of 2.5m unique words, with each word being represented by a 300-dimension vector. A subset of this model was extracted for each of the three textbooks that are analyzed in this paper.

2.3 Recommending Vocabulary to Study

In this section, we introduce a method of creating a map of the vocabulary for each textbook. A word in the vocabulary is represented as a node, and the weight of edges between each pair of words represents the semantic context distance between the 2 words. The vector representations of vocabulary that were obtained from the word2vec model were used to measure the context distance by calculating the cosine distance between the vectors of two words. As a number of the items in the vocabulary list contain multiple words that would result in an equal number of vectors which could not be compared using standard techniques. A method of representing these words as a single vector is necessary, and the sum total of the word embedding vectors as proposed in Lau & Baldwin (2016) was used to represent these items. The map contains a large number of edges, and an optimal sub-map of the strongest relations was extracted by the minimum spanning tree algorithm as described in Flanagan et al (2019). This results in a map that links words by the closest similarity while keeping the number of edges in the map at a minimum. The distribution of the similarity context distance of all of the edges of the maps before and after optimization are shown in Fig. 1 along with the mean and standard deviation in Table 2.

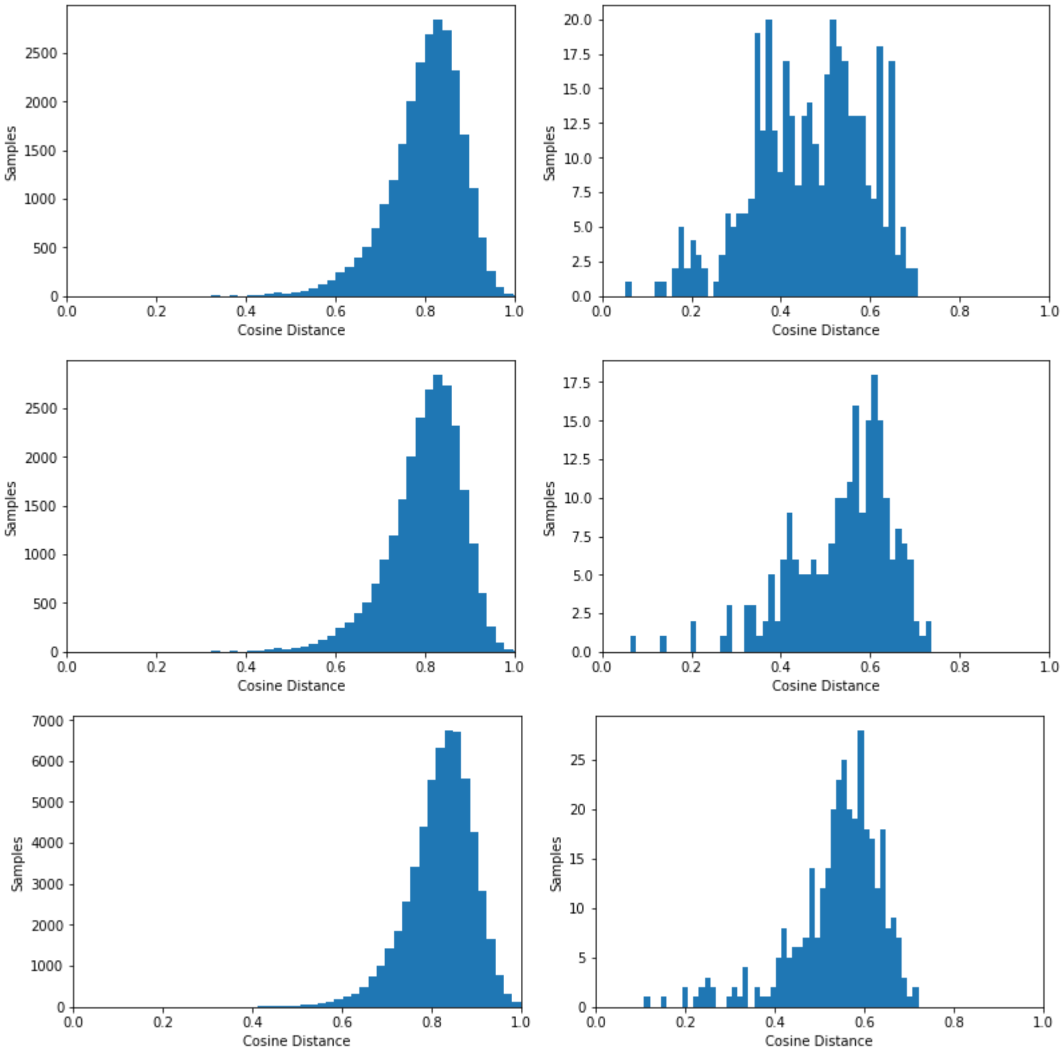

Figure 1. Distribution of semantic distance of all word relations (left) and of optimized word relations (right) for all year 1 (top) to year 3 (bottom) of junior high school.

Table 2
An overview of the three different levels of EFL learning materials that were targeted in this paper.

	Year 1		Year 2		Year 3	
	Mean	SD	Mean	SD	Mean	SD
All	0.7884	0.0911	0.7996	0.0824	0.8193	0.0740
Optimized	0.4645	0.1244	0.5384	0.1116	0.5415	0.0996

The distribution of the semantic context distance of vocabulary in the textbook for year 1 has a lower mean and greater standard deviation than higher years. This suggests that the range of vocabulary broadens as the target proficiency increases. Also, the mean distance of the optimized map is substantially lower than that of all relations, indicating that the relation of the vocabulary has greater semantic context similarity.

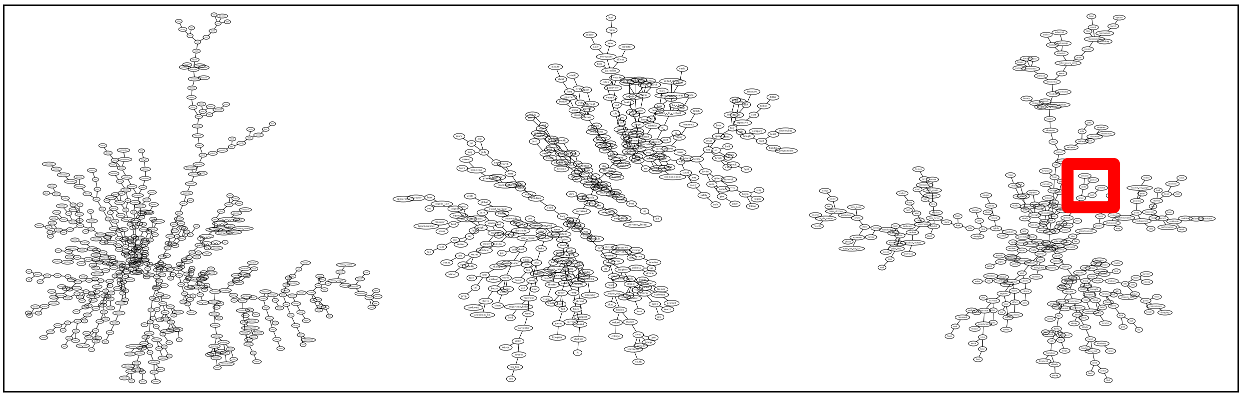

Figure 2. Overview of semantic vocabulary learning maps for year 1 (left) to year 3 (right) of junior high school.

An overview of the maps generated for three different textbooks using the proposed method based on cosine distance are shown in Fig. 1. Some of the branches in the map are long and are usually related to specific themes that are covered in the textbook, whereas shorter branches mostly cover specific language concepts.

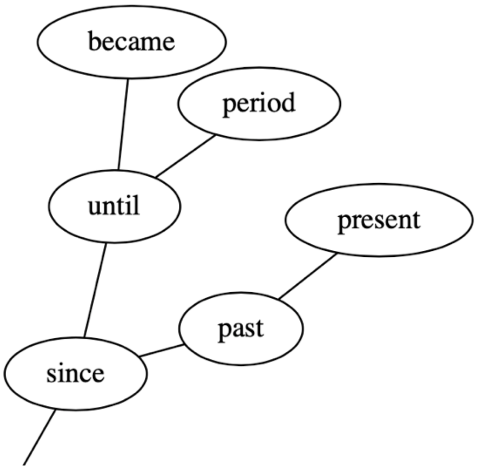

Figure 3. Details of a branch representing vocabulary that relate to time concepts that are studied in third year junior high school textbook.

An example of the types of vocabulary paths that the map can recommend a student to learn are shown in Fig. 3, which contains a detailed view of the red highlighted branch in Fig 2. It is made up of words that are related to the concept of time that students learning in the third year of junior high school. One sub-branch "past" and "present" represents the tense time concept, while other parts are related to time relations, such as: since, until, and period. An example of recommendations that could be made using this map would be if a learner has just studied the vocabulary "since" the system would then recommend that the learner next studies the words "until" and "past". These vocabularies have the same semantic context of time in common with the parent word "since". If the learner chooses to then study "past", the system could then recommend that they should study the word "present" afterwards. The weights of the edges that are connected to the current node could be used to rank the recommendation, giving higher importance to vocabulary that have a close semantic context to previously learnt words.

In addition to recommending vocabulary based on the semantic context, recommendations could should also be based on the difficult of the word. As described in section 2.1, we extracted an ordered list of vocabulary from the textbooks based on the first occurrence of words. On the assumption that the textbooks would introduce increasingly difficult words as study progresses, this list can be analyzed to recommend vocabulary based on the difference of the position. For example the tense branch occurs in the following sequence within the textbook: "since" → "present" → "past", which suggests that a learning should study the vocabulary in this order.

3. Conclusion

In this paper, we propose a method to automatically generate a map of the semantic context relations of vocabulary, and define how it can be used to recommend vocabulary items that a learner could study based on their learning history. The relation of the vocabulary items was determined by measuring the cosine distance between two words in a word2vec corpus that was trained on the sentence contexts of the English Wikipedia corpus. We applied the method to three textbooks that are frequently used in Japanese junior high schools when studying English as a foreign language and examined parts of the maps that were generated as case studies. The maps will be introduced to a knowledge mapping system (Flanagan et al., 2019) that is being deployed in Japanese K-12 institutions to support the study of English.

In future work, we plan to formally evaluate the use of the recommendation system in Japanese K-12 English as a foreign language class.

Acknowledgement

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innovation Program on AI and Big Data 18102059-0.

References

Azab, M., Hokamp, C., & Mihalcea, R. (2015). Using word semantics to assist English as a second language learners. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 116-120).

Borodkin, K., Kenett, Y. N., Faust, M., & Mashal, N. (2016). When pumpkin is closer to onion than to squash: The structure of the second language lexicon. *Cognition*, *156*, 60-70.

Flanagan, B., Akçapınar, G., Majumdar, R., & Ogata, H. (2018). Automatic Generation of Contents Models for Digital Learning Materials. *Proceedings of the 26th International Conference on Computers in Education (ICCE2018)* (pp. 804-806).

Flanagan, B., Majumdar, R., Akçapınar, G., Wang, J., & Ogata, H. (2019). Knowledge Map Creation for Modeling Learning Behaviors in Digital Learning Environments, *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge (LAK19)*.

Lau, J. H., & Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *ACL 2016* (pp. 78-86).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Nation, I., & Hunston, S. (2013). Learning words from context. In *Learning Vocabulary in Another Language* (Cambridge Applied Linguistics, pp. 348-388). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139858656.010

Wolter, B. (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. *Applied Linguistics*, 27, 4, 741–747.