Course Concepts: How Readable Are They for ESL Learners?

Yo Ehara
Tokyo Gakugei University
ehara@u-gakugei.ac.jp

ABSTRACT

Massive open online courses (MOOCs) are online courses for multiple learners with different backgrounds, including English-as-a-second-language (ESL) learners. In a MOOC, course concepts are important for diverse learners to grasp what they can learn in the course and its prerequisite knowledge. Previous studies have explored methods to automatically extract concepts from course videos or identify prerequisite concepts in a course. However, as a concept typically consists of several words, it could be difficult for ESL learners to understand what a concept means if they do not know the words in the concept. For example, for "geospatial data," many of them may need an additional explanation of what "geospatial" means in addition to the explanation of the concept. This paper extensively analyzes the readability of MOOC concepts using an openly-available manually-annotated MOOC-concept dataset on computer science and economics and a vocabulary test result dataset of ESL learners with different English skills. We found that the percentage of concepts for which an ESL learner is likely to know all the words is only 25.8% in computer science. In economics, the value is 56.5%. This implies that ESL learners usually require additional vocabulary explanations to understand MOOC concepts. We also show qualitative analyses and that almost half of the concepts are unreadable to ESL learners.

Keywords

Course Concepts, Readability, Second Language Learners

1. INTRODUCTION

Massive open online courses (MOOCs) are online lecture courses intended for use by a large number of learners with different backgrounds, including English-as-a-second-language (ESL) learners. In MOOCs, students learn numerous knowledge concepts, or course concepts, some of which are taught in the course, whereas others are prerequisites of the course. Course concepts are important because learners "with different backgrounds can grasp the essence of the course" [5]. Previous studies have focused on extracting course concepts automatically from course video recordings [5] or identifying prerequisite concepts of a course [5]. However, MOOC learners also include ESL learners. How much additional effort is required to ensure that ESL learners understand the concepts taught in the course? No previous study has extensively investigated this research question, which we address in this paper.

For example, consider the concept of "big data." ESL learners who are willing to listen to an English course usually know both "big" and "data" because both "big" and "data" are high-frequency words on the general corpus; therefore, they are likely to have been mastered by the learners in their previous English studies. In this case, the teacher only needs to explain what "big data" is. Therefore, the effort to teach this concept to ESL learners is almost the same as that to native English speakers. In contrast, when considering the concept of "geospatial data," it is possible that many ESL learners do not know the meaning of the word "geospatial". "Geospatial" is a specialized word that is rare in general corpora. While native English speakers may only need an explanation of "geospatial data," an ESL learner may need an additional explanation of what "geospatial" means, such as "something related to locations and maps." No previous studies have extensively studied the difference in the effort required to teach a concept to native English speakers and ESL learners.

This study estimates how much additional effort is required when teaching concepts to ESL learners using an openly available manually checked MOOC concept list dataset. Specifically, we estimate which words learners are likely to know the meaning of by using a machine-learning method that takes vocabulary test results and the frequency of the general corpus as features. We experimented with manually annotated concept datasets from online courses in computer science and economics. The experiment showed that 60% of the concepts consisted of two English words, and approximately half of the concepts are not readable to almost all learners in the learner vocabulary dataset that we employed.

2. DATASETS

Unlike academic wordlists and specialized terminology extraction studies and their datasets, MOOC concepts refer to the specific knowledge taught in MOOCs. One of the openly available English course concept datasets manually verified

Y. Ehara. Course concepts: How readable are they for esl learners? In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 426–429, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115742

Table 1: Bigram concepts with largest difference between two words in computer science.

Words	Difference	Mean Prob.
right subtree	0.655	0.215
left subtree	0.642	0.212
block tridiagonal	0.581	0.190

Table 2: Bigram concepts with largest difference between two words in economics.

Words	Difference	Mean Prob.
OCO order	0.621	0.215
rediscounted rate	0.602	0.210
salesforce management	0.596	0.209

is that of [5], which is a collection of concepts taken from eight computer science and economics courses on Coursera, one of the most popular English MOOCs. While larger MOOC concept datasets are available in subsequent studies, namely MOOCCube and MOOCCubeX, they are taken from XuetangX, which mainly consists of Chinese courses. Hence, throughout the paper, we use the dataset by [5].

To answer our research question, we also need a dataset from which we can obtain what kinds of words ESL learners know. Since MOOCs are intended to offer courses for many learners with diverse backgrounds over the Web, the ESL learners of the dataset are also preferred to have been collected on the Web. Few datasets meet this criterion because, in most ESL datasets, ESL learners are classroom students of a school; hence, they are not diverse. One such dataset is [1], in which 100 ESL learners answer 100 vocabulary questions. The learners of this dataset were collected using crowdsourcing; hence, they have more diverse backgrounds than classroom learners.

3. EXPERIMENTS

The dataset of [5] contains eight Coursera computer science and economics courses, including their transcripts. Human annotators manually annotated whether k-grams in the transcripts were course concepts or not. In computer science, in total, the dataset has 4,096 concepts; nearly 60% of them consist of two words (bigrams), 18% one word (unigrams), and 22% three words (trigrams). In economics, in total, the dataset has 3,652 concepts; nearly 66% of them consist of bigrams, 10% unigrams, and 24% trigrams.

We also built a classifier that predicts how likely a word is to be known to a learner. To this end, we used the learner

Table 3: Bigram concepts with smallest difference between two words in computer science.

Words	Difference	Mean Prob.
learning rule	9.94×10e-5	0.645
thread programming	3.74×10e-4	0.468
network traffic	4.38×10e-4	0.614

Table 4: Bigram concepts with smallest difference between two words in economics.

Words	Difference	Mean Prob.
federal agency	3.16×10e-6	0.643
quantitative easing	3.69×10e-5	0.426
domestic cresit	1.04×10e-4	0.659

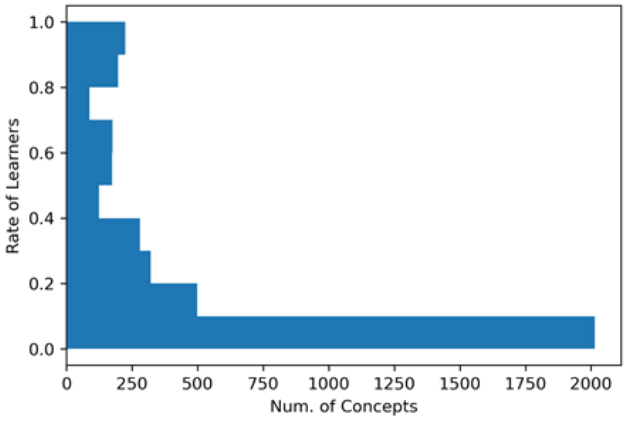

Figure 1: Histogram of Concepts Known to Learners in computer science.

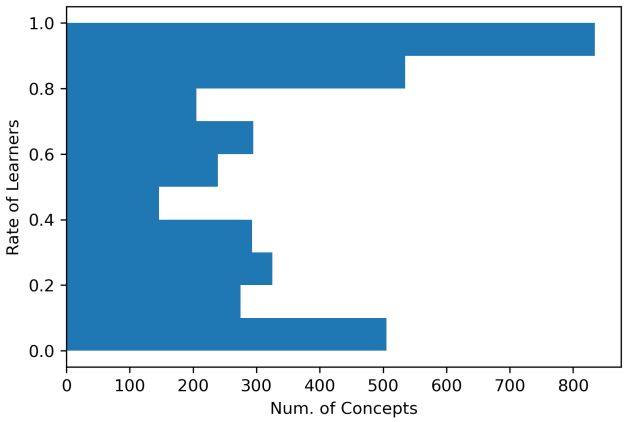

Figure 2: Histogram of Concepts Known to Learners in economics.

vocabulary test dataset of [1]. We built a machine-learning classifier that, given a learner and a word, classifies whether the learner knows the word. Following [1], we used one-hot vectors for learner features and word frequencies taken from the British National Corpus (BNC) and Contemporary Corpus of American English (CoCA) as features; both corpora are general corpora frequently used for teaching ESL learners. In the [1] dataset, the learners' English skills in this dataset are diverse while the test-takers are mainly Japanese because the dataset was built using a Japanese crowdsourcing service called Lancers. The dataset consists of 100 ESL learners; those who do not know even the basic words "computer" and "science" are unlikely to be willing to learn computer science in English, we omitted such lowly skilled learners, resulting in 94 learners. For classification, we used logistic regression because it was previously applied to their dataset and was reported to have high accuracy in measuring ESL learners' readability [3]. For the 10,000 responses (100 learners for 100 vocabulary questions) of the dataset, we first split the data into 9,800 for training data and 200 for test data. The logistic regression was highly accurate as it achieved 86.1% accuracy on the test data in the [1] dataset, whereas the chance rate was 60.3%.

We then applied our classifier to the MOOC concepts. Specifically, for each learner in the dataset, we obtained the probability that the learner knows the word for each word in a concept. By simply taking the product of the probability values of all words in a concept, we obtained the probability that the learner knows the concept: for example, when learner A knows "big" and "data" with probabilities of 0.9 and 0.8, respectively, the probability that learner A knows "big data" is 0.72. Then, if the probability of a concept is equal to or greater than 0.5, we considered that the learner knows the concept.

In computer science, on average, an ESL learner knows 1,058 concepts, which amounts to only 25.8% of the 4,096 concepts, implying that an ESL learner needs an explanation to understand some word(s) in the concept. Figure 1 is a histogram showing what percentage of ESL learners the concepts are known to. We can see that almost half of the concepts are known to less than 10% learners.

In economics, situations are quite different from those in computer science. On average, an ESL learner knows 2,065 concepts, which amounts to only 56.5% of the 3,652 concepts, implying that an ESL learner needs an explanation to understand some word(s) in the concept. Figure 2 is a histogram showing what percentage of ESL learners the concepts are known to. We can see that almost 500 of the concepts are known to less than 10% learners, whereas almost 800 concepts are known to more than 90% learners.

We then focus on the average ESL learner in the dataset and see the concepts that may require special attention when teaching second language learners. To this end, we focus on the bigram concepts and see the difference in the probability known to ESL learners between the two words of which each concept consists. Whereas native-speaker learners know both words and simply need to learn what the concept as a whole means, in addition, ESL learners need to learn what the word in the concept means if the learner does not know the meaning of a word in the concept. What is particularly unintuitive is that one word of the concept is easy for learners, whereas the other(s) is/are not. In this case, the words constituting the concept may seem easy to native-speaker teachers because one of the words is easy. However, as the other word(s) is/are not, such concepts can be confusing to ESL learners. Hence, we list up these words in the following paragraphs.

In computer science, Table 1 shows the bigram concepts with the largest difference in the mean probability known to the average learner between the two words in the concepts, and Table 3 shows the concepts with the smallest difference. We can see that the words particularly difficult for the average learner were "subtree" and "tridiagonal."

In economics, Table 2 shows the bigram concepts with the largest difference in the mean probability known to the average learner between the two words in the concepts, and Table 4 shows the concepts with the smallest difference. We can see that the words particularly difficult for the average learner were "OCO" and "rediscounted."

4. RELATED WORK AND DISCUSSION

In this study, we used concept data from an English MOOC. On the other hand, if the language is not limited to English, a study on MOOCs includes data from a large MOOC in Chinese in [7]. Conceptual information is expensive for teachers to tag, so the study [8] helps teachers by automatically assigning conceptual information. Such research will eventually be used to recommend courses for MOOCs [9].

However, these studies have not paid particular attention to the common case of MOOC participants being second language learners. As for the readability of second language learners, there are mainly two approaches to collecting the dataset for experiments.

One approach is to collect data from language teachers. In this approach, language teachers teaching second language learners read each text in the dataset and label the difficulty. Particularly, the task for automatically assessing the readability of texts is called automatic readability assessment (ARA) and has been studied extensively in [6, 4]. The strength of this approach is that we can easily obtain one gold label for each text. The weakness of this approach is that the quality of the annotations heavily depends on the expertise of the language teachers.

In contrast, another approach is to collect data from language learners themselves. English learners cannot directly annotate what texts are difficult for them. However, unlike the method of having English teachers annotate the texts, this method can obtain information directly from the English learners. Therefore, it is not affected by the noise of what kind of students the English learners have taught in the past. In this approach, data from language learners taking a vocabulary test consisting of short sentences is available to the public [1]. This study also followed this approach. Especially, [2] investigates the readability of scientific abstracts.

5. CONCLUSIONS

To conclude, we made preliminary analyses of the readability of MOOC concepts to ESL learners. Importantly, Figure 1 shows that, for nearly 2,000 concepts of the 4,096 ones, ESL learners also need an explanation of the words used in the concept to understand the explanation of the concept. According to the Figure 2, this situation is relaxed in the field of economy, but still, about 500 concepts out of 3652 concepts, or 13.7%, are not understood by ESL learners.

These results indicate that if ESL learners could know the meaning of the basic words used in the concepts before taking these courses, their understanding of the courses might be greatly improved. To this end, future work includes personalized support systems that automatically explain the words in the concepts.

6. ACKNOWLEDGMENTS

This work was supported by JST ACT-X Grant Number JPMJAX2006, Japan. We used the ABCI infrastructure of AIST. We appreciate the anonymous reviewers' valuable comments.

7. REFERENCES

[1] Y. Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.

[2] Y. Ehara. Semantically adjusting word frequency for estimating word difficulty from unbalanced corpora. In *Companion Proc. of LAK*, 2020.

[3] Y. Ehara. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In *Proc. of Educational Data Mining (short paper)*, 2022.

[4] M. Martinc, S. Pollak, and M. Robnik-Šikonja. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179, Apr. 2021.

[5] L. Pan, X. Wang, C. Li, J. Li, and J. Tang. Course concept extraction in MOOCs via embedding-based graph propagation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 875–884, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing.

[6] S. Vajjala and I. Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[7] J. Yu, G. Luo, T. Xiao, Q. Zhong, Y. Wang, W. Feng, J. Luo, C. Wang, L. Hou, J. Li, Z. Liu, and J. Tang. MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020.

[8] J. Yu, C. Wang, G. Luo, L. Hou, J. Li, J. Tang, M. Huang, and Z. Liu. ExpanRL: Hierarchical Reinforcement Learning for Course Concept Expansion in MOOCs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 770–780, Suzhou, China, Feb. 2020. Association for Computational Linguistics.

[9] H. Zhang, X. Shen, B. Yi, W. Wang, and Y. Feng. KGAN: Knowledge Grouping Aggregation Network for course recommendation in MOOCs. *Expert Systems with Applications*, 211:118344, Jan. 2023.