Measuring Similarity between Manual Course Concepts and ChatGPT-generated Course Concepts

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

ABSTRACT

ChatGPT is a state-of-the-art language model that facilitates natural language interaction, enabling users to acquire textual responses to their inquiries. The model's ability to generate answers with a human-like quality has been reported. While the model's natural language responses can be evaluated by human experts in the field through thorough reading, assessing its structured responses, such as lists, can prove challenging even for experts. This study compares an openly accessible, manually validated list of "course concepts," or knowledge concepts taught in courses, to the concept lists generated by ChatGPT. Course concepts assist learners in deciding which courses to take by distinguishing what is taught in courses from what is considered prerequisites. Our experimental results indicate that only 22% to 33% of the concept lists produced by ChatGPT were included in the manually validated list of 4,096 concepts in computer science courses, suggesting that these concept lists require manual adjustments for practical use. Notably, when ChatGPT generates a concept list for non-native English speakers, the overlap increases, whereas the language used for querying the model has a minimal impact. Additionally, we conducted a qualitative analysis of the concepts generated but not present in the manual list.

Keywords

Language Models, Course Concepts, Computer Science

1. INTRODUCTION

ChatGPT is a state-of-the-art natural language processing (NLP)-based artificial intelligence (AI) chatbot system released by OpenAI on November 30, 2022, and can answer any question you enter in a dialogue format. For example, in education, it can be used to answer simple code generation and short essays, and early reports say that the system has surprisingly excellent quality in many tasks. However, its answers may contain factual or logical errors. For codes, essays, and other textual items longer than a sentence, a teacher or expert can read them and find errors. However, for those with simpler structures, such as lists, it is difficult for even teachers to detect errors.

In Massively Open Online Courses (MOOCs), typically, learners can freely choose which courses to take. The concepts taught in MOOCs are important for learners to decide which courses to take because the concepts help learners understand what they should learn in the course and what are prerequisites. Since it is time-consuming for a teacher to create a list of concepts in a course, methods were previously proposed to generate the list directly from course transcripts or course materials [1]. However, even while using these, we still need to collect transcribed courses and materials.

If we ask ChatGPT to "tell us about concepts that will be important in computer science learning," will it be possible to produce a high-quality list of concepts automatically? To determine this, it is necessary to evaluate the quality of ChatGPT's output, but human teachers are not good at evaluating list formats.

2. DATASETS

In this study, we need a list of manually identified concepts. If the concept list is based on use within a specific school or region, it may have been based on assumptions about the educational system of that school or region. For example, a list of concepts from a particular university might include the name of the computer systems of that university, or what is learned in high school in the country where the university resides might be treated as something known by all learners and not included in the list. Since it is undesirable to use such a biased list for evaluation, we used concept lists for MOOCs.

[1] offers an openly available MOOC concept list. Their goal was to create concept lists automatically from course transcripts. For this purpose, the concept lists were manually extracted from course transcripts of eight computer science courses on Coursera, a well-known website for MOOCs in English. The list has 4,096 concepts in total. Subsequent works by [1], such as MOOCCube [2] and MOOCCubeX [3], contain much larger lists of concepts. However, these data are Chinese concepts based on XuetangX, a MOOC system whose courses are predominantly in Chinese. Although English translations of these data sets are also provided, we did not use them in this study because they raise the question of whether the list of concepts used in Chinese courses

Y. Ehara. Measuring similarity between manual course concepts and chatgpt-generated course concepts. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 474–476, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115758

Table 1: Overlap Rate with Manual List.

Lang. for Prompt	Gen. for what students	Overlap rate
English	(not specified)	0.222
English	Japanese	0.315
Japanese	Japanese	0.336

corresponds directly to the list of concepts in English.

Many studies have created academic wordlists or lists of technical terms in English, but it is difficult to strictly define "academic" or technical terms in these studies. Unlike these studies, in this study, we focuse more specifically on course concepts that learners actually learn in online computer lectures. Thus, words such as "introduction," which are academic in the sense that they are often used in academic papers but do not express specific concepts in a field, are excluded from the concepts.

3. EXPERIMENTS

In this study, three lists were created for three use cases, assuming a variety of students. First is the use case in which we want to list the concepts that English-speaking students need to learn when studying computer science in English. Second is the use case in which we want to list the concepts that Non-Native English Speaker (NNS) students need to learn when studying computer science in English. Last is the use case in which we want to do this for NNS students by asking ChatGPT in the students' native language instead of English. Japanese was chosen as the language other than English.

The list was generated using ChatGPT. Input to a language model such as ChatGPT to generate something is called a "prompt". For example, the following prompt was used to ask ChatGPT to list concepts that Japanese students would need in an English computer science course.

- "List 40 concepts that Japanese students need to learn when they study computer science in English online courses on computer science."

The reason for specifying 40 concepts is the length limitation of the answers. However, ChatGPT can also ask questions related to the previous question. Therefore, the following additional prompt will generate a list of 40 concepts that are different from the previous one: "List another set of 40 concepts that differs from the previous one." By entering additional prompts like this, a total of about 120 responses can be obtained for each use case. For English-speaking students, we used the prompt in which the word "Japanese" was simply removed from the aforementioned prompt.

4. RESULTS

Table 1 lists the "overlap rate" as the percentage of concepts generated by ChatGPT included in the list of manually confirmed concepts. Note that the list of manually identified concepts is more comprehensive, since the list of manually identified concepts is 4,096, while only about 120 are generated by ChatGPT. "Lang. used for prompt" indicates the language used for the question, and "Gen. for what students" describes the adjective before the word "student" in the prompt example above, such as "Japanese". As shown in Table 1, the highest percentage was generated for Japanese students in Japanese. Conversely, there was no significant difference in the overlap rate for the languages used, i.e., "Lang. used for prompt".

Table 2: Generated but not in Manual List.

relational database, normalization, bus, decidability, transaction, huffman coding, primitive recursive function, float, array, private key, run-length encoding, captcha, object-oriented programming, turing machine, rest, digital signature, loop, arithmetic coding, brute force

The reason for this is future work. Qualitatively, when the type of student was not specified, the generated concepts tended to have more abbreviations for practical content than for theoretical content. Also, specifying "Japanese students" may have implicitly specified generating concepts for university students because studying abroad is more popular among university students. Table 2 shows the words that were not included in the manually generated list for Japanese students in Japanese. Thus, qualitatively, all words appear to represent "concepts". The reasons why these words were not included are also covered in our future work. Notably, the human-made concept list used in this study was made by annotating words that appeared in the actual spoken lectures. Thus, it could be possible that these concepts, although relevant to the courses, tend to be related but are actually not frequently spoken during courses.

5. DISCUSSION

In this study, the course concept lists generated by ChatGPT were compared to manually generated concept lists. The resulting overlap values between ChatGPT-generated course concepts and manually-created course concepts were low. However, the generated course concept lists do not appear to be low quality since almost all of them represent some concepts of informatics, although the overlap values were low.

Hence, the main result of this study, the overlap values, are limited in its generalizability. The low overlap values could possibly indicate that ChatGPT and other language models cannot generate high-quality course concept lists. However, there are other possibilities, as follows.

First, the generated human course concept list may not be exhaustive, while we employed seemingly the most exhaustive manually-created course concept list to the best of our knowledge. In this case, the overlap values would be low regardless of the performance of ChatGPT in generating course concept lists.

It is also important to note that there is a five-year gap between 2017 when the human-handled course concept list was built [1], and 2022, when ChatGPT was introduced. Hence, it is possible that the low overlap values do not indicate ChatGPT's limited capabilities but rather that the trends

in informatics have changed over the past five years.

Furthermore, ChatGPT itself is updated daily. Therefore, if the latest version of ChatGPT is used, it is likely that the overlap values may be improved without any special efforts.

6. CONCLUSIONS

ChatGPT is known for its ability to generate text in a variety of formats. Text fluency can be more easily evaluated by native speakers by reading, while evaluation of list format is difficult for humans. In this study, we evaluated the properties of lecture concept lists, which are important for learners to select lectures, by having ChatGPT generate them. Compared to an exhaustive human list of 4096 lecture concepts in the field of computer science, only up to 33% of the list generated by ChatGPT was included in the human list of lecture concepts. This indicates that the focus of ChatGPT as a lecture concept list is different from the focus of human beings when creating a lecture concept list.

If the number of lecture concept lists is small, there will naturally be lecture concepts that are not included in the list, even if they were created manually. This time, we used the most comprehensive list of lecture concepts in a single field that has been created manually. On the other hand, the list was biased toward one field, computer science. Future work will be to evaluate the generation of lecture concept lists by ChatGPT for other fields as well.

7. 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.

8. REFERENCES

[1] 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.

[2] 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*, pages 3135–3142, Online, July 2020. Association for Computational Linguistics.

[3] J. Yu, Y. Wang, Q. Zhong, G. Luo, Y. Mao, K. Sun, W. Feng, W. Xu, S. Cao, K. Zeng, Z. Yao, L. Hou, Y. Lin, P. Li, J. Zhou, B. Xu, J. Li, J. Tang, and M. Sun. Mooccubex: A large knowledge-centered repository for adaptive learning in moocs. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, CIKM '21, page 4643–4652, New York, NY, USA, 2021. Association for Computing Machinery.