Augmenting Case Based Learning With Dynamic Language Models

Conrad Czejdo
Fayetteville State University
cczejdo1@broncos.uncfsu.edu

Sambit Bhattacharya, IEEE Senior Member
Fayetteville State University
sbhattac@uncfsu.edu

***Abstract*—This paper describes a novel supporting tool for the case-based learning (CBL). Recent advances in deep-learning based language models (LMs) have enabled highly dynamic interactivity in dialog services and story generation. We leverage the progress in modelling language to develop a technologically augmented CBL pedagogy which we analyze with a standardized assessment. Our assessment shows reasonable case interactivity, low rates of factual inaccuracy, and no inappropriate machine-sourced responses. We also compare our assessment results across the case categories of Ethics, Chemistry, Biology, and Medicine, but find no statistically significant differences. In summary, we develop a framework for analyzing the ability of LMs to augment CBL, apply this framework to the GPT-3 LM, and discuss some of the challenges and potential solutions to ensuring proper usage in the classroom environment.**

***Index Terms*—Pedagogical Techniques, Deep Learning (DL), Natural Language Processing (NLP), Case Based Learning (CBL), Language Model (LM)**

I. Introduction

Case-Based Learning (CBL) is an established teaching method which has seen extensive use across many fields [1] [2] [3] [4] . By correlating examples with real-world or real-world-like situations, CBL seeks to provide learners with greater context and an easier transition to practical application. With regards to educational efficacy over other leading pedagogies, studies show mixed results [4]. However, the evidence is clearer that students show increased enjoyment, collaborativity, and ability to connect to practice. Furthermore, educational professionals advocate for CBL since it allows for increased reflection and creativity during the decision-making process, which is especially crucial for certain professions such as Medicine and Law. Current educational processes are being further impacted due to the recent explosion of Machine Learning (ML). The general problem of ML is that of modelling data in such a way that the input and outputs of the model match a training set while maintaining generalizability to examples beyond those specifically used during training. Deep Learning (DL) is a particular subset of ML which has garnered an exceptional amount of attention due to a vast library of extensions, ability to model vast complexity, and (in most cases) a straightforward method of increasing performance by adding more data. Education technology companies, like Khan Academy, Knewton, and ASSISTments, have used DL on datasets to model educational data [5] [6]. One example of such modelling is predicting the future assessment results of a student based on problems they have solved (correctly or incorrectly) thus far. The trained models were capable of not only predicting problem set ordering, but also in providing insights which helped optimize prerequisite pathways. Recent work in the area, however, has shown that educational datasets are not necessarily better analyzed with DL versus more traditional ML approaches (e.g. logistic regression, support vector machines, etc.) [7]. Although the need for DL in certain segments of educational data analysis is debated, there are areas in which it has defined strengths. One of these areas is language modelling. The problem definition of language modelling is generally predicting text from contextual text (e.g. predicting future text from previous text or predicting what fills in a blank from surrounding text) [9]. Related to LMs is translation, where providing input text from one language should result in output text in another language. DuoLingo, for example, used DL to help define a set of possible outputs for their language learning app [8]. The application of DL has been especially instrumental in the development of very large scale LMs [9] which are trained on enormous corpuses of text to extract patterns of writing. With the vast amount of information these LMs store, they have been shown to be capable of reasonably answering questions with only a few – and in some cases zero – examples. Furthermore, these LMs have been used in generating cohesive and interactive stories from text-based user input [10]. Inspired by these uses, we contribute an analysis of the current applicability of LMs towards an interactive CBL system by conducting a study of its responses to a standardized assessment.

II. Methodology

This study's primary goal was to assess the ability of language models to augment case studies with automated interactive components. We conducted an assessment of the most pertinent challenges related to incorporating language models, namely of their ability to respond in appropriate, coherent, and satisfactory ways.

A. Data

We used 25 case study abstracts (e.g. Fig.1) from the National Center for Case Study Teaching in Science [13]. Four prompt categories were present: Medicine (7), Ethics (7), Biology (6), and Chemistry (4). We filtered each abstract text to exclude specific mentions of prerequisites, school level, and

SoutheastCon 2021 | 978-1-6654-0379-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/SOUTHEASTCON45413.2021.9401836

978-1-6654-0379-5/21/$31.00 ©2021 IEEE

> Since its first recorded appearance in 1996, Tasmanian Devil Facial Tumor Disease (DFTD) has wiped out an estimated 70 percent of the Tasmanian devil population. Sci……Students also consider possible methods for containing and eliminating DFTD.

Fig. 1. Abridged example of case study abstract from the National Center for Case Study Teaching in Science [13].

additional learning resources (e.g. videos and powerpoints). The average length of each prompt was 630 ± 170 characters, with no significant differences between categories.

B. Language Model

We initially attempted to utilize previously available LMs, such as the largest publicly released GPT-2 [11] model as well as smaller versions of the currently available GPT-3 (model names: *ada*, *babbage*, and *curie* – from smallest to largest). OpenAI also released models which were further fine-tuned for the specific task of responding to instructions, named *curie-instruct* and *davinci-instruct*. It was only with this largest (175 billion parameter) and latest fine-tuned model (davinci-instruct) that we found output which was reasonable (rather than copying the prompt word for word, prematurely ending, or erroneous text) in the context of our experiment. GPT-3 uses a token-based system for output, where each token is a chunk which can contain more than a letter, but is usually less than a full word (equivalent to around 4 characters). To sample these tokens, we set the nucleus sampling and temperature parameters to (0.7). The max number of tokens per response was set to (200). We also set the frequency and presence penalties to (1.0). The combination of these parameters provided us with an invaluable reduction in repeat responses while maintaining sensible and consistent answers.

C. Assessment Errors

The following is a hierarchical classification of indicators which we believe encompass important aspects to test an LM based system for CBL. A count of each error occurrence is kept for analysis.

1) Fatal Errors: If these are encountered, the query which caused the error is removed. The reason we define these as fatal errors is that if the offending query is not removed, then the model tends to continue to reply with the same response.

- *Repeating Response:* If a response is an exact replica of a previous response or prompt.
- *Premature End:* If the response is blank.

2) Significant Errors: These errors require supervisor intervention, and can potentially result in significant questions about the ability of the system to participate in a school environment.

- *Inappropriate Response:* Broadly defined as language not appropriate for school grounds; significant examples include sexually charged statements or slurs. Identified as an important area of OpenAI's research focus [9].
- *Factually Incorrect:* A statement that goes against current scientific knowledge, and that could result in confusion and significant time investment by the teacher to correct. Note: If the same factual inaccuracy is repeated throughout the case interaction, we still only count it as a single error.

3) Coherence Errors: These errors make users question the polish of the system, but do not necessarily require intervention. Note: We allow the error value count to be split based on sub-queries (e.g. if 2 out of 3 options from a query which asks for 3 responses provide a Low Quality response, then the number of Low Quality errors increase by 0.67)

- *Low Quality:* These responses repeat the question in a different form, or have significant grammatical and/or syntactical errors.
- *Numerical Error:* When asking for a specific number of sentences or options, and the response provides a different number of them.
- *Poor Analogy:* Analogies are important learning tools, but it can also be difficult to make a sensible one. As such, we define this as a separate error.

D. Standard Assessment

For each abstract in our dataset, we apply a standard assessment. In Fig.2 we define the standard assessment with the queries we provide to the model. We also specify that the interactive story ("text adventure") should be aimed at middle schoolers to provide an easier target for the LM, and also to indicate we are looking for school-appropriate language. We ask for three introductions to test for how diverse the LM responses can be. We also ask the LM to generate three questions since intermittent questioning is an important component of CBL.

III. Results and Discussion

In this section, we present the results of our standard assessment applied to LM augmented case study abstracts. Furthermore, we discuss possible reasons as well as solutions to the presented error tallies.

A. Aggregated Results

Fig. 3 shows the average errors present per assessment. The most predominant errors are the "coherence errors" (Low Quality Response, Numerical Error, or Poor Analogy), and students are very likely to encounter them during their interaction with the LM. Repeating text seemed to be another common problem, however this issue is a bit more worrisome since it is a "fatal error" which requires the removal of the offending query. Offsetting this concern a bit, the other fatal error, Premature End, seemed to occur with relative

Prompt: Abstract: <Abstract Text Here>
Q1: Please give 3 potential two-sentence introductions (numbered 1,2,3) for a middle school appropriate interactive text adventure based on the preceding abstract.
Q2: Write two more sentences about option 1.
Q3: What are the learning objectives of this story?
Q4: Provide and explain an analogy for <Passage Concept>.
Q5: Explain <Passage Concept>.
Q6-Q8: <Interaction With Case Story>
Q9: Write 3 questions about the story so far based on the abstract.
Q10: Write 3 one-sentence explanations to the previously asked questions.

Fig. 2. Ten standard queries asked of the LM each interaction with a new prompt. Initially, the prompt and Q1 is provided. If Q1 is removed due to causing a fatal error, then we remove Q1 but keep the prompt and move on to asking Q2. All italicized text is entered as a query to the model. <Abstract Text Here> is filled with the abstract text for the case study. <Passage Concept> is a pertinent concept found in either the produced story or abstract (e.g. evolution bottleneck, Devil Facial Tumor Disease, etc.). The text which replaces <Interaction with Case Story> is more open-ended and meant to mimic the type of interaction a student would have with the story presented by the LM so far.

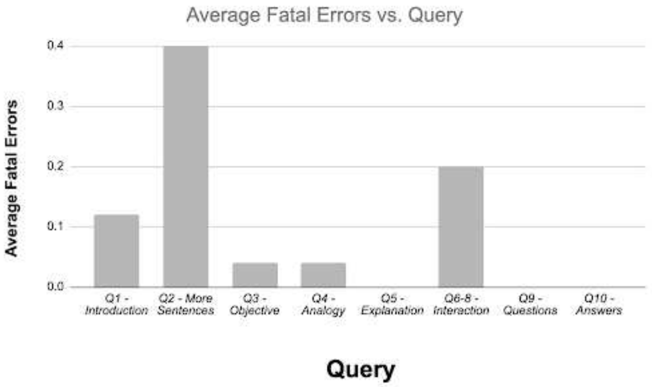

Fig. 3. Average errors present in each case assessment.

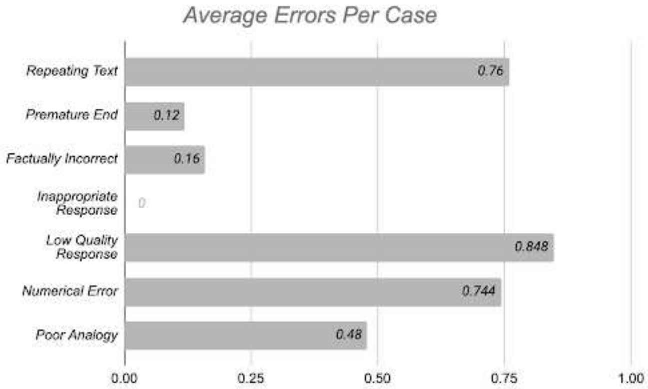

Fig. 4. Average fatal errors present in each case assessment at multiple query points.

infrequency. Fig. 4 shows that one of the major causes of the presence of fatal errors was the second query, which asked to continue one of the story introductions presented by the LM from the first query. Since continuing a story without defining the "fork in the road" decision seems to work better in comparison (as seen in Q6-8), teachers should be advised to follow a more linear form of interaction, at least with current LM. "Significant errors" seemed to be of least worry, as the "Factually Incorrect" error only occurred 16% of the time, while machine-sourced Inappropriate Responses occurred zero times. Concerns regarding these errors would probably be more significant in an actual classroom, where students have the ability to interact however they wish with the system. A possible solution could be limiting the queries students can make, or implementing an inappropriate language detector to alert the teacher if someone is misusing the system. Unfortunately, although specific queries can be fine-tuned to reduce these errors, the bulk of the improvement will have to be done by fundamental improvements in the LM itself – whether that is further increasing training data, increasing trainable parameters, or more sophisticated DL techniques.

B. Case Type Distinctions

We further analyzed the average errors present by case type (Fig.5). We used a one-way ANOVA test for each distribution. We found no statistically significant ($p<.05$) differences related to Numerical Errors, Premature Ends, Factual Incorrectness, Inappropriate Text, Low Quality Responses, or Repeating Text. Although our case category analysis was inconclusive at our sample size, increasing our testing may help uncover more subtle differences between categories.

C. Highlighted Examples

Although LMs have their difficulties in answering some questions, as is evidenced by Fig.6, in other cases they can provide good insights. In Figure 7, the abstract prompt was a case about selecting candidates for a job. The LM was able to utilize this information to generate a set of candidates with descriptions. Later on, the LM defined the salary each candidate was looking for, and finally said whether a good candidate was picked!

IV. Conclusion

In this research, we utilized LMs to augment CBL, and conducted an analysis using a standardized assessment. Our standardized assessment showed no LM-sourced inappropriate responses, meaning that filtering user input could work as well as filtering LM output. We found that even the most advanced LMs still have a noteworthy number of coherence and repeating text errors. However, most of the repeating errors can be avoided with a careful choice of query at the beginning of the prompt. This underlines the specific needs of the queries for LMs, but other examples (Fig. 6 and 7) show great potential in the novel interactivity LMs provide.

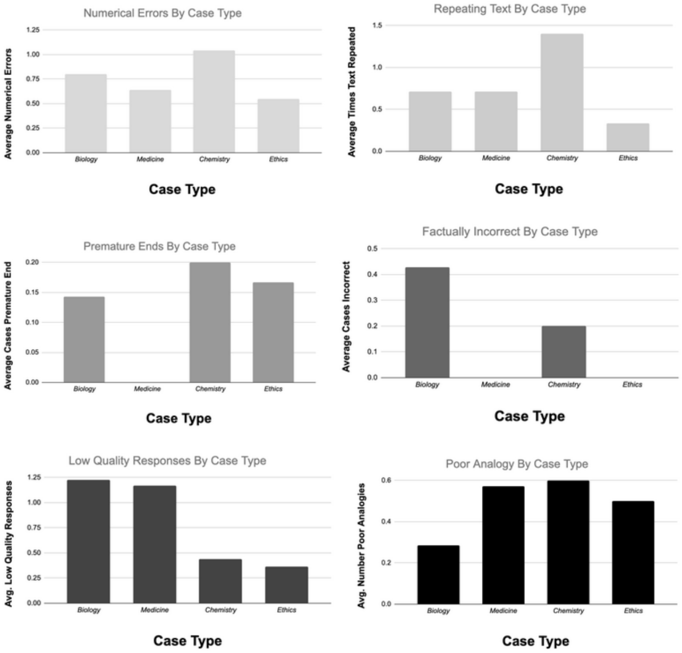

Fig. 5. Average errors present in each case assessment, by case type.

Provide and explain an analogy for greenhouse gas emissions. Greenhouse gas emissions are like the fuel that makes our cars go.

Provide and explain an analogy for immunity. An analogy for immunity is like a shield that protects you from viruses.

Fig. 6. Example of poor (top) and good (bottom) analogies generated by LM.

Provide a short description of each candidate. Candidate 1: This candidate has a strong background in environmental science and would be a good fit for the position… … Candidate 5: While there isn't much information about Candidate 5 available, what we do know is that they have experience working with large teams and seem like they would make a great addition to your team!

Fig. 7. Example of good story continuation, where new information is produced by the LM.

ACKNOWLEDGMENT

This article is based upon work supported by the National Science Foundation under Grant No. 1818694. We would also like to thank OpenAI for supplying free access to their GPT-3 class of language models.

REFERENCES

[1] S. Gade and S. Chari (2013). ”Case-based learning in endocrine physiology: an approach toward self-directed learning and the development of soft skills in medical students,” Advances in physiology education, 37(4), 356–360.

[2] L. Kantar and A. Massouh (2015). ”Case-based learning: What traditional curricula fail to teach,” Nurse Education Today, 35(8), e8-e14.

[3] M. Baeten, F. Dochy, K. Struyven (2013). ”Enhancing students' approaches to learning: the added value of gradually implementing case-based learning.” Eur J Psychol Educ, 28, 315–336.

[4] J. E. Thistlethwaite, D. Davies, S. Ekeocha, J. Kidd, C. MacDougall, et al. (2012). ”The effectiveness of case-based learning in health professional education,” A BEME systematic review: BEME Guide No. 23. Medical Teacher, 34(6), e421–e444.

[5] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, et al. (2015). ”Deep Knowledge Tracing,” In Proceedings of the 29th Conference on Advances in Neural Information Processing Systems. 505–513.

[6] C. Perrotta and N. Selwyn (2019). ”Deep learning goes to school: toward a relational understanding of AI in education.” Learning, Media and Technology, 45(3), 251–269.

[7] T. Doleck, D.J. Lemay, R.B. Basnet, and P. Bazelais (2020). Predictive analytics in education: a comparison of deep learning frameworks. Educ Inf Technol 25, 1951–1963.

[8] S. Mayhew, K. Bicknell, C. Brust, B. McDowell, W. Monroe, and B. Settles (2020). ”Simultaneous Translation and Paraphrase for Language Education,” Proceedings of the Fourth Workshop on Neural Generation and Translation. https://doi.org/10.18653/v1/2020.ngt-1.28

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al. (2020). ”Language Models are Few-Shot Learners.” arXiv:2005.14165. unpublished.

[10] Latitude. (n.d.) AI Dungeon. https://play.aidungeon.io/

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, “Language Models are Unsupervised Multitask Learners.” (2019). unpublished.

[12] OpenAI (n.d.) OpenAI API Beta. https://beta.openai.com/

[13] “National Center for Case Study Teaching in Science.” National Center for Case Study Teaching in Science (NCCSTS), sciencecases.lib.buffalo.edu/.