Exploring Artificial Jabbering For Automatic Text Comprehension Question Generation

Tim Steuer[0000−0002−3141−712X], Anna Filighera[0000−0001−5519−9959], and Christoph Rensing

Technische Universität Darmstadt, Hesse, Germany
{tim.steuer,anna.filighera,christoph.rensing}@kom.tu-darmstadt.de

Abstract. Many educational texts lack comprehension questions and authoring them consumes time and money. Thus, in this article, we ask ourselves to what extent artificial jabbering text generation systems can be used to generate textbook comprehension questions. Novel machine learning-based text generation systems jabber on a wide variety of topics with deceptively good performance. To expose the generated texts as such, one often has to understand the actual topic the systems jabbers about. Hence, confronting learners with generated texts may cause them to question their level of knowledge. We built a novel prototype that generates comprehension questions given arbitrary textbook passages. We discuss the strengths and weaknesses of the prototype quantitatively and qualitatively. While our prototype is not perfect, we provide evidence that such systems have great potential as question generators and identify the most promising starting points may leading to (semi) automated generators that support textbook authors and self-studying.

Keywords: Text comprehension · Language models · Automatic question generation · Educational technology

1 Motivation

Reading, alongside direct verbal communication, is one of the most prevalent forms of learning. For every new subject, we encounter in our educational careers, highly motivated educators publish textbooks to help us understand. Even after we finish our formal education, the modern knowledge society is based on lifelong informal learning in which learners in the absence of teachers, also often devote themselves to textual learning resources. In both, the formal and informal scenarios only gaining surface-level understanding is likely not enough. If we e.g. study a physics or history textbook to pass an exam deeper understanding of the topic is crucial. However, reading is difficult and to deeply comprehend a text, passive consumption is insufficient [25,7].

Instead, readers need to actively reflect the information provided in the text to reach a deep understanding [25,7]. A well-explored method to actively engage readers is posing questioning about what they have read [25,1]. Yet, posing good questions consumes time and money and thus many texts encountered

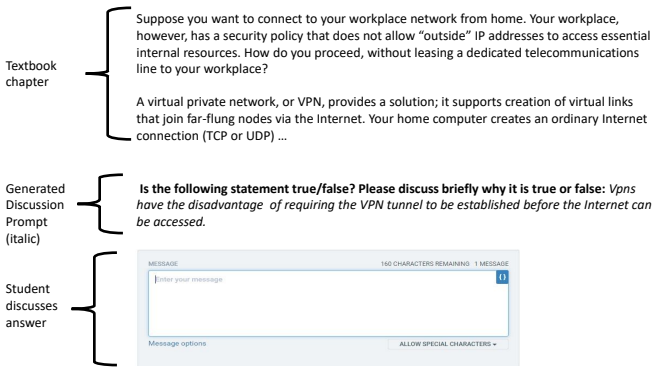

Fig. 1. Example usage of the proposed system.

by learners either contain only a few questions at the end of a chapter or lack questions.

Educational automatic question generation investigates approaches to generate meaningful questions about texts automatically, reducing the necessity for manually generated questions. It hereby relies either on machine learning-based approaches that excel in question variety and expressiveness but pose mostly factual questions [6] or on rule-based approaches that lack expressiveness and variety [32] but have limited capability to pose comprehension questions depending on their purpose (e.g. [17]).

This article investigates a novel machine learning-based question generation approach seeking to generate comprehension questions with a high variety and expressiveness. We hereby rely on two main ideas. First, research in the educational domain has investigated learning from errors [19] indicating that explaining why a statement or solution is faulty may foster learning, conceptual understanding, and far transfer [10]. Second, we rely on the artificial jabbering of state-of-the-art neural text generators that are capable of extrapolating a given text with high structural consistency and in a way that often looks deceptively real for humans. We seek to explore whether this jabbering can be conditioned in such a way that it generates erroneous examples from textbook paragraphs. Presented with such a statement, learners need to justify if a given statement is true or false (see Figure 1). This work comprises three main contributions:

1. We present the idea of leveraging artificial jabbering for automatic text comprehension question generation and introduce a prototypical generator.
2. We provide a quantitative and qualitative evaluation of the strengths and weaknesses of such an approach.
3. We distill the main challenges for future work based on an in-depth error analysis of our prototypical generator.

2 Related Work

2.1 Learning from erroneous examples

When learning with erroneous examples, students are confronted with a task and its faulty solution and have to explain why it is wrong (e.g. [30]). The underlying theoretical assumptions are that erroneous examples induce a cognitive conflict in students and thus support conceptual change [24] e.g. by pointing out typical misconceptions [29]. It has been shown that erroneous examples are beneficial for learning in a variety of domains such as mathematics [10], computer science [4] or medicine [14]. Also, learners confronted with erroneous examples especially improve deeper measures of learning such as conceptual understanding and far transfer [24]. However, some studies have found that erroneous examples only foster learning when learners receive enough feedback [30,14] and have sufficient prior knowledge [30].

2.2 Neural Text and Question Generation

With the rise of high capacity machine-learning models, language generation has shifted towards pretraining [27]. Trained on huge datasets, these models provide state-of-the-art results on a wide variety of natural language generation tasks [23,5] such as dialog response generation tasks [22] or abstractive summarization tasks [26]. Novel models like GPT-2 [23] are capable of extrapolating a given text with high structural consistency and in a way that looks deceptively real for humans. They copy the given text's writing style and compose texts which seem to make sense at first glance. Fine-tuning the model even increased the humanness of the generated texts [28]. Research in the credibility of such generated texts found that hand-picked generated news texts were found to be credible around 66% of the time, even when the model was not fine-tuned on news articles [28]. Another study found that human raters could detect generated texts in 71.4% of the cases with two raters often disagreeing if the text is fake or not [13]. These findings started a debate in the natural language generation community if the model's generation capabilities are to easy to misuse and therefore the models should not be released anymore [28]. Furthermore, such models are able to generate poems [16] and to rewrite stories to incorporate counterfactual events [21]. Besides of these open text generation models, special models for question generation exist. They evolved from baseline sequence to sequence architectures [6] into several advanced neural architectures (e.g. [33,5]) with different facets such as taking the desired answers into account [34] or being difficulty-aware [8]. Although these systems work well in the general case they are mainly focusing on the generation of factual questions [6,35,20]. Thus, although their expressiveness and domain independence is impressive, the educational domain still most often uses template-based generators [15]. These template-based approaches are often able to generate comprehension questions but lack expressiveness and rely on expert rules limiting them to a specific purpose in a specific domain.

3 An Experimental Automatic Erroneous Example Generator

To experiment with the idea of using artificial jabbering for improving text comprehension, we propose the following text generation task. The input is a text passage of a learning resource from an arbitrary domain, having a length of 500-1000 words as this has been used in psychological studies that found text accompanying questions to be helpful [1,31]. The output is a generated text comprehension question about the given text passage, asking learners to explain why a given statement is true or false. We aim to generate high-quality questions of good grammaticality, containing educational valuable claims and having the right difficulty for discussion. Some technical challenges are inherent in the described task. Every approach must tackle discussion candidate selection as this determines what the main subject of the generated text will be. Also, every approach must provide the neural text generator with a *conditioning context* to ensure that the generated text is in the intended domain. Finally, every approach must render the actual text with some sort of open domain generator. These subtasks are active fields of research and a huge variety of possible approaches with different strengths and weaknesses exists. Yet, our first aim is to evaluate the general viability of such an approach. Thus, we do not experiment with different combinations of sub-components but our generator relies on well-tested domain-independent general-purpose algorithms for the different subtasks (see Figure 2).

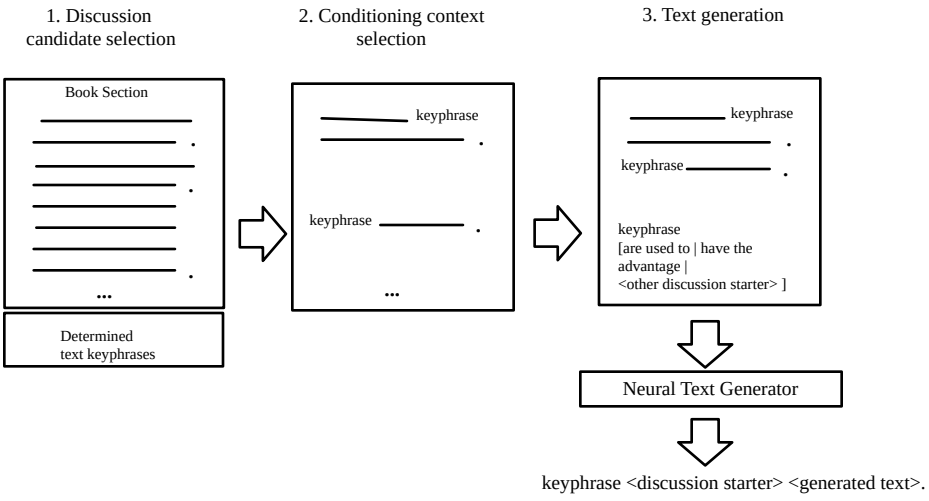

Fig. 2. Architecture of the automatic text comprehension question generator. The final output is a justication statement that is combined with a prompt to form the actual text comprehension question.

First, for the *discussion candidate* selection, we make the simplifying assumption that good discussion candidates are the concepts that are characteristic of the text. To understand why this assumption is simplified consider a text about Newtonian physics where a few sentences discuss the common misconception that heavier objects fall faster than lighter objects. This discussion is unlikely to involve any special keywords and thus will not be selected as input to the generator. Yet, it might be very fruitful to generate erroneous examples based on these misconceptions. However, to test our general idea of generating erroneous examples the simplification should be sufficient because we might select fewer inputs but the one we select should be important. Furthermore, this assumption allows us to rely on state-of-the-art keyphrase extraction algorithms. Considering that the inputs are texts from a variety of domains, the keyphrase selection step needs to be unsupervised and relatively robust to domain changes. Therefore, we apply the YAKE keyphrase extraction algorithm [3] which has been shown to perform consistently on a large variety of different datasets and domains [2]. Stopwords are removed before running keyphrase extraction and the algorithm's configured windows size is two.

Second, for selecting the *conditioning context*, a short text that already comprises statements about the subject is needed. Suppose the discussion subject is "Thermal Equilibrium" in a text about physics. For the generator to produce interesting statements it must receive sentences from the text, discussing thermal equilibria. Thus, we extract up to three sentences in the text comprising the keyphrase, by sentence tokenizing the text [1] and concatenating sentences containing the keyphrase.

Third, we need to generate a justification statement as the core for the text comprehension question. We use the pretrained GPT-2 774M[2] parameter model and apply it similar to Radford et al. [23] by using plain text for the model conditioning. The plain text starts with the sentences from the *conditioning context* and to generate the actual justification statement, a discussion starter is appended. It begins with the pluralized discussion subject followed by a predefined phrase allowing us to choose the type of justification statement the model will generate. For instance, let "Thermal Equilibrium" be our discussion subject, our to be completed discussion starter may be "Thermal equilibria are defined as" or "Thermal equilibria can be used to" depending on the type of faulty statement we aim for. The resulting plain text is given to GPT-2 for completion. To prevent the model from sampling degenerated text, we apply nucleus sampling [12] with top-p=0.55 and restrict the output length to 70 words. Finally, we extract the justification statement from the generated text and combine it with a generic prompt to discuss it, resulting in the final text comprehension question. Note that we do not know, if the generated question is actually comprising a true or false justification statement.

[1] using NLTK-3.4.5
[2] https://github.com/openai/gpt-2

4 Research Question and Methodology

4.1 Research Question

We evaluate our generation approach on educational texts from a variety of domains focusing on the following research question:

RQ: To what extent are we able to generate useful text comprehension statements in a variety of domains given short textbook passages?

Looking at the related work, a fraction of the generated statements should already be usable without any adjustments, while many other statements need adjustment. We conduct a quantitative evaluation and qualitative evaluation. Our procedure includes a quantitative expert survey, a qualitative error analysis to determine useful error categories and a qualitative analysis of the already usable results to better describe their features.

4.2 Methodology

Quantitatively, a total of 120 text comprehension questions coming from ten educational texts are annotated by ten domain experts who have been teaching at least one university lecture in a similar domain. Texts are equally distributed across five different domains: Computer Science, Machine Learning, Networking, Physics and Psychology. Twelve text comprehension questions are generated for every text. They are based on three extracted discussion candidates and four different discussion starters, of which we hypothesized that they represent intermediate or deep questions according to Graesser et. al [9]. The discussion starters are: "X has the disadvantage", "X has the advantage", "X is defined as" and "X is used to" where X is the discussion candidate. This Every question is rated by two experts who first read the educational text that was used to generate the question and then rate it on five five-point Likert items regarding grammatical correctness, relatedness to the source material, factual knowledge involved when answering the question, conceptual knowledge involved when answering the questions and overall usefulness for learning. Before annotating every expert saw a short definition of every scale, clarifying their meaning. Additionally, experts can provide qualitative remarks for every question through a free-text field. For the quantitative analysis the ratings where averaged across experts.

We use the quantitatively collected data to guide our qualitative analysis of the research questions. To carry out our in-depth error analysis, we consider a statement useless for learning if it scores lower than three on the usefulness scale. This choice was made after qualitatively reviewing a number of examples. We use the inductive qualitative content analysis [18] to deduce meaningful error categories for the statements and to categorize the statements accordingly. Our search for meaningful error categories is hereby guided by the given task formulation and its sub-components. Furthermore, the useful generated statements (usefulness$\geq$ 3) are analyzed. We look at the effects of the different discussion starters and how they influence the knowledge involved in answering the generated questions.

5 Results

5.1 Quantitative Overview

The quantitative survey results indicate that many of the statements generated are of good grammar, are connected to the text but are only slightly useful for learning (see Figure 3). Furthermore, most questions involve some factual knowledge and deeper comprehension, yet both scores vary greatly. Breaking down the different rating scores by domain or discussion starter does not revealed no large differences. By looking at various examples of different ratings (see Table 1) we found that a usefulness score of three or larger is indicative of some pedagogical value. With minor changes, such questions could be answered and discussed by experts, although their discussion is probably often not the perfect learning opportunity. In total, 39 of the 120 statements have a usefulness rating of 3 or larger (32.5%), in contrast to 81 statements rated lower (67.5%).

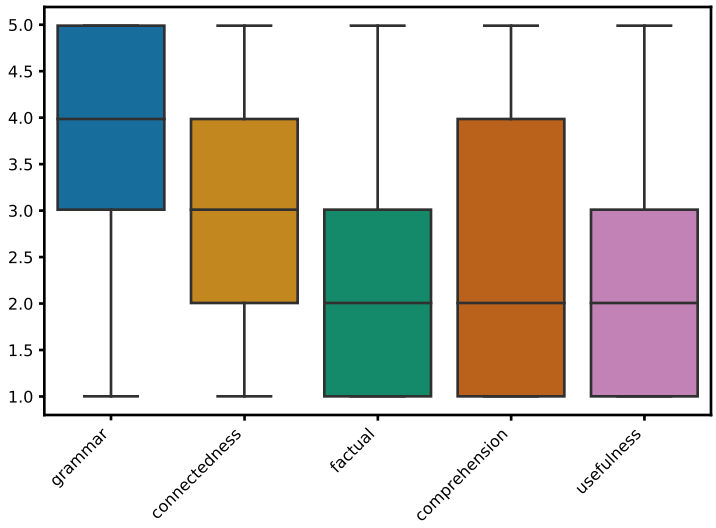

Fig. 3. Overview of the quantitative ratings for the generated statements without any human filtering. Scores are between 1 and 5 where 5 is the best achievable rating. The whiskers indicate 1.5 Interquartile range and the black bar is the median.

5.2 Qualitative Error Analysis

While conducting the qualitative error analysis, the following main error categories where deduced. *Keyword inappropriate* means that the discussion candidate was not appropriate for the text because the keyword extraction algorithm

Table 1. Examples of differently rated generated statements (higher = better).

Usefulness ranking	**Example statement**
1	Fastest possible machines have the disadvantage of being more expensive to build and maintain.
2	Prior knowledges have the advantage that they are easy to measure and easy to measure the causal role of.
3	Knowledge bases can be used to test the performance of models, and to improve the performance of inference engines.
4	Von neumann architectures have the advantage of being able to process a wide range of instructions at the same time, making them highly scalable.
5	Vpns have the disadvantage of being difficult to set up and maintain, and they can be compromised by bad actors.

selected a misleading or very general key term. *Keyword incomplete* means that the discussion candidate would be good if it would comprise additional terms. For example, in physics, the discussion candidate sometimes was "Equilibrium" instead of "Thermal Equilibrium". *Platitude* means that the generated statement was a generic platitude and thus not helpful. *Hardly discussable* means that the statement was either too vague or too convoluted therefore making it hard to write a good justification. Finally *too easy* means that the students could answer by just relying on common sense.

Table 2. The different error categories and their distribution

inappropriate keyword	**incomplete keyword**	**platitude**	**hardly discussable**	**statement too easy**
43	6	9	11	12

The distribution of the different error categories can be seen is heavily skewed towards keyword errors (see Table 2). The two keyword-based errors account for 49 or roughly 60% of the errors. Furthermore, statements generated by faulty keyword selection mostly have a usefulness rating of one. The other error categories are almost equally distributed and are most often rated with a usefulness score of two. The *platitude* case mostly comes from unnaturally combining a discussion candidate with a discussion starter resulting in very generic completion of the sentence inside the generator. For instance, if the generator has to complete the sentence "Classical conditionings have the disadvantage ..." it continues with " ...of being costly and slow to develop". The remaining error categories have no clear cause.

Besides the error analysis, annotators left some remarks about the erroneous statements. Two annotators remarked on various occasions that the first part of the sentence (discussion candidate + discussion starter) is incomprehensible and

thus the whole statement is worthless. One annotator remarked that there are missing words in the keyword leading to a bad rating for the statement. The keyword was for example "knowledge" instead of "knowledge base". Furthermore, one annotator remarked that the statement has not enough discussion potential. Those comments are in line with our deduced error categories for keyword errors.

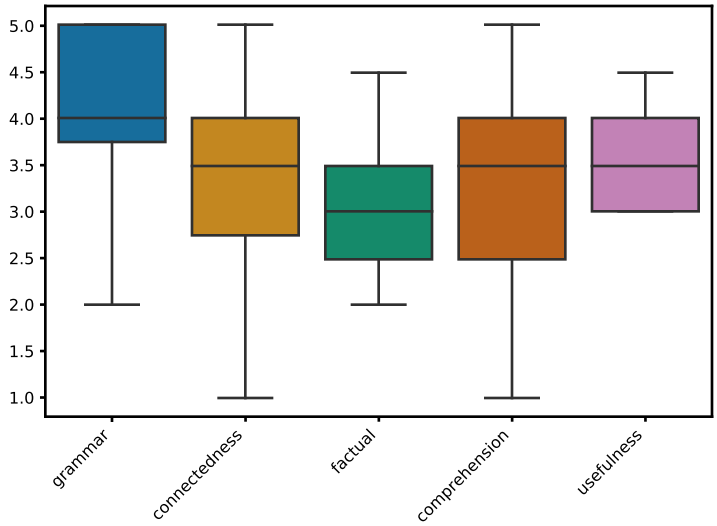

Fig. 4. Overview of the quantitative ratings for the generated statements with an usefulness rating larger or equal three. Scores are between 1 and 5 where 5 is the best achievable rating. The whiskers indicate 1.5 Interquartile range and the black bar is the median.

5.3 Quality Characteristics of the Useful Statements

The 39 statements with a usefulness rating larger three also score well in the other factors (see Figure 4). Especially, the involved factual knowledge and deeper comprehension clearly increase. Reviewing the generated statements reveals that the generated statements are not simply a paraphrase of a fact stated in the text. Thus, learners answering the corresponding question cannot simply do a keyword spotting but need to think about the actual content of the text. Furthermore, the generated statements adequately use technical terminology. Moreover, the different discussion starters play an important role as they lead to different types of statements. When generating with the *definition* starter, the generator rephrases the definition of a discussion candidate in "its own words". As a result, these definitions often lack important aspects or contain faulty claims (see Table 3). Thus, to explain why the definition is wrong, learners have to compare and contrast their previous knowledge with the generated definition. The *usage* starter

leads to statements that force learners to transfer the knowledge learnt into new situations (see Table 3). The usage that is described in the generated statements is normally not mentioned in the text, but can often be deduced by the knowledge provided in the text. The *advantage and disadvantage* discussion starter requires learners to think about the discussion candidate but also about similar concepts and solutions and to compare them (see Table 3). Otherwise, learners cannot tell if the stated advantage or disadvantage is one that is specific to the discussed concept.

Table 3. Highly rated examples of different types of statements resulting from different discussion starter.

Discussion starter	**Domain**	**Example statement**
Definition	machine learning	Knowledge bases are defined as data structures that store knowledge and act as a long-term memory.
(Dis)advantage	psychology	Conditionings have the advantage of being simple and universal, which makes them ideal for studies of complex behavior.
Usage	networking	Hosts can be used to forward packets between hosts.

Finally, one annotator provided qualitative remarks for the good statements as well. This includes remarks that the generated statements are helpful but often could be improved by using different discussion starters depending on the domain (e.g. speaking of the advantage of a physical concept is odd). Also, it was highlighted that the statements cannot simply be answered by copying information from the text and that thinking about the *definition* discussion starter sometimes resulted in the annotator checking a textbook to refresh some rusty knowledge.

6 Discussion and Future Work

Concerning our research question, we can say that roughly a third of the statements have some educational value. This is in line with the related work that reports between 29% and 66% deceptively real statements [28, 13]. Yet, even lower numbers of valuable statements can be beneficial. If we do not generate questions directly for the reader, but for textbook authors for further review, it can be a source of creative ideas and may reduce the authoring effort. In particular, such systems could be combined with question ranking approaches similar to Heilman et. al [11] to only recommend the most promising candidates.

Furthermore, there is more to our research question than just this quantitative view and looking at our qualitative results reveals interesting characteristics of the well-generated statements. First, they are not the typical factual Wh-questions that ask for a simple fact or connection directly stated in the text. Therefore, they often need a deeper understanding of the subject matter to be

answered correctly. While this can be a benefit, we have to keep in mind that our annotators were experts and thus drawing connections between the text inherent knowledge and previously learned subject knowledge might be too difficult for some learners as also remarked by the annotators. Second, depending on the used discussion starter, we can generate different kinds of useful questions. Our four different discussion starters generate questions requiring three different types of thinking. Depending on the discussion starter, the text comprehension questions involve comparison with previous knowledge, transfer of learned knowledge to new situations or implicit differentiation from similar concepts. An encouraging result, because it shows that the generator's expressiveness can be harnessed to create different types of tasks. Moreover it provides evidence for the remark of the annotators, that the questions in some domains could be improved by using different discussion starters and that this is a worthwhile direction for future research. Third, although we work with a variety of domains and input text from different authors we were able to generate some valuable questions in every domain. Furthermore, the distribution of the different quality scores did not change much from domain to domain. Hence, our approach seems, at least to a degree, domain-independent. Yet, as currently only a third of the generated statements are usable this should be reevaluated as soon as the general quality of the statements becomes better because it might be a trade-off between domain-independence and statement quality. In summary, our qualitative analysis of the well-generated questions provided evidence for their adaptability through different discussion starters and that they are well suited for text comprehension below the surface level when learners have to think not only about facts but also have to integrate knowledge.

Our error analysis allowed us to identify why we fail to generate interesting questions. The five different error categories are promising starting points for future work. Most often, the approach failed because the keyword extraction step did not find a meaningful discussion candidate or extracted only parts of it. This is not surprising as our goal was to test the general idea without fine-tuning any of the intermediate steps. General-purpose keyword extraction is similar but not identical to discussion candidate extraction. Hence, future work might explore specific educational keyword extraction algorithms and their effect on the generation approach. We assume that a fine-tuned educational keyword extraction algorithm will yield much more valuable statements if adaptable to different domains. Furthermore, as discussed in the results section the platitude errors can be alleviated by not combining discussion starters and discussion candidates in an odd manner. Future work should, therefore, investigate the optimal use of discussion starters taking into account different domains and discussion candidates. Finally, we have the *hardly discussable* and *statement too easy* error categories. While no clear cause of these errors could be identified, we assume that a fine-tuning of the neural generator with discussion specific texts would reduce these types of errors. The related work has already shown that fine-tuning neural generators yields performance gains [28]. Yet, one has to be careful not to lose some of the expressiveness of the current model. Thus, future work might explore the

relation between fine-tuning for the generation of justification statements and the change in the expressiveness of the model. One important issue thereby is that fine-tuning should allow the model to generate more erroneous statements comprising typical misconceptions of learners as these are particularly beneficial for learning [29]. Besides of the actual generation process, feedback to learners' answers is crucial and should be explored further [29].

Finally, we would like to point out some limitations of the current study. First, our goal was to explore the general idea of the generation of questions and not finding the optimal approach. Hence, this work only provides a lower bound for the quality of the generated questions and other state-of-the-art keyword extraction algorithms and language generators might yield better performance. We think our work is valuable nevertheless, as it demonstrates a working prototype, the key advantages of such a prototype and provides a strong baseline for future work. Furthermore, while other combinations of algorithms might yield better performance, we provided an in-depth error analysis to inform the research community on what to focus on. Third, while asking experts to score the statements is often used in research it is unclear if the experts' opinion correlates with the actual perception of students. However, we assume that the experts' assessment agrees with learners' at least in tendency and that this is sufficient to assess the basic generation idea. Fourth, we are aware that qualitative analysis is always to some degree subjective. Yet, we believe that for complex novel approaches such, as the one presented, it is an important and often neglected way of collecting valuable data about the inner workings of the approach.

To conclude, artificial jabbering of neural language models has the potential to foster text comprehension as it has unique strengths not present in other neural question generators. The initial implementation in this work may be used as a tool for authors, providing them with ideas about what they could ask students. However, it is too error-prone to interact directly with learners and we provided valuable pointers to improve this in future work.

References

1. Anderson, R.C., Biddle, W.B.: On asking people questions about what they are reading. Psychology of Learning and Motivation - Advances in Research and Theory **9**(C), 89–132 (1975). https://doi.org/10.1016/S0079-7421(08)60269-8
2. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., Jatowt, A.: Yake! keyword extraction from single documents using multiple local features. Information Sciences **509**, 257–289 (2020)
3. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., Jatowt, A.: Yake! collection-independent automatic keyword extractor. In: European Conference on Information Retrieval. pp. 806–810. Springer (2018)
4. Chen, X., Mitrovic, T., Mathews, M.: Do novices and advanced students benefit from erroneous examples differently. In: Proceedings of 24th International Conference on Computers in Education (2016)
5. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.W.: Unified language model pre-training for natural language understanding and

generation. In: Advances in Neural Information Processing Systems. pp. 13042–13054 (2019)
6. Du, X., Shao, J., Cardie, C.: Learning to Ask: Neural Question Generation for Reading Comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 8, pp. 1342–1352. Association for Computational Linguistics, Stroudsburg, PA, USA (2017). https://doi.org/10.18653/v1/P17-1123, http://aclweb.org/anthology/P17-1123
7. Duke, N.K., Pearson, P.D.: Effective practices for developing reading comprehension. Journal of education **189**(1-2), 107–122 (2009)
8. Gao, Y., Bing, L., Chen, W., Lyu, M.R., King, I.: Difficulty controllable generation of reading comprehension questions. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 4968–4974. AAAI Press (2019)
9. Graesser, A., Rus, V., Cai, Z.: Question classification schemes. In: Proc. of the Workshop on Question Generation. pp. 10–17 (2008)
10. Große, C.S., Renkl, A.: Finding and fixing errors in worked examples: Can this foster learning outcomes? Learning and instruction **17**(6), 612–634 (2007)
11. Heilman, M., Smith, N.A.: Good question! statistical ranking for question generation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 609–617. Association for Computational Linguistics (2010)
12. Holtzman, A., Buys, J., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019)
13. Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.: Human and automatic detection of generated text. arXiv preprint arXiv:1911.00650 (2019)
14. Kopp, V., Stark, R., Fischer, M.R.: Fostering diagnostic knowledge through computer-supported, case-based worked examples: effects of erroneous examples and feedback. Medical education **42**(8), 823–829 (2008)
15. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education pp. 1–84 (2019)
16. Liao, Y., Wang, Y., Liu, Q., Jiang, X.: Gpt-based generation for classical chinese poetry. arXiv preprint arXiv:1907.00151 (2019)
17. Liu, M., Calvo, R.A., Rus, V.: G-asks: An intelligent automatic question generation system for academic writing support. Dialogue & Discourse **3**(2), 101–124 (2012)
18. Mayring, P.: Qualitative content analysis. A companion to qualitative research **1**, 159–176 (2004)
19. Ohlsson, S.: Learning from performance errors. Psychological review **103**(2), 241 (1996)
20. Pan, L., Lei, W., Chua, T.S., Kan, M.Y.: Recent advances in neural question generation. arXiv preprint arXiv:1905.08949 (2019)
21. Qin, L., Bosselut, A., Holtzman, A., Bhagavatula, C., Clark, E., Choi, Y.: Counterfactual story reasoning and generation. arXiv preprint arXiv:1909.04076 (2019)
22. Qin, L., Galley, M., Brockett, C., Liu, X., Gao, X., Dolan, B., Choi, Y., Gao, J.: Conversing by reading: Contentful neural conversation with on-demand machine reading. arXiv preprint arXiv:1906.02738 (2019)
23. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
24. Richey, J.E., Andres-Bray, J.M.L., Mogessie, M., Scruggs, R., Andres, J.M., Star, J.R., Baker, R.S., McLaren, B.M.: More confusion and frustration, better learning: The impact of erroneous examples. Computers & Education **139**, 173–190 (2019)

25. Rouet, J.F., Vidal-Abarca, E.: Mining for meaning: Cognitive effects of inserted questions in learning from scientific text. The psychology of science text comprehension pp. 417–436 (2002)
26. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368 (2017)
27. See, A., Pappu, A., Saxena, R., Yerukola, A., Manning, C.D.: Do massively pretrained language models make better storytellers? In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). pp. 843–861 (2019)
28. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Wang, J.: Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203 (2019)
29. Tsovaltzi, D., McLaren, B.M., Melis, E., Meyer, A.K.: Erroneous examples: effects on learning fractions in a web-based setting. International Journal of Technology Enhanced Learning **4**(3-4), 191–230 (2012)
30. Tsovaltzi, D., Melis, E., McLaren, B.M., Meyer, A.K., Dietrich, M., Goguadze, G.: Learning from erroneous examples: when and how do students benefit from them? In: European Conference on Technology Enhanced Learning. pp. 357–373. Springer (2010)
31. Watts, G.H., Anderson, R.C.: Effects of three types of inserted questions on learning from prose. Journal of Educational Psychology **62**(5), 387 (1971)
32. Willis, A., Davis, G., Ruan, S., Manoharan, L., Landay, J., Brunskill, E.: Key phrase extraction for generating educational question-answer pairs. In: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale. pp. 1–10 (2019)
33. Zhang, S., Bansal, M.: Addressing semantic drift in question generation for semi-supervised question answering. arXiv preprint arXiv:1909.06356 (2019)
34. Zhao, Y., Ni, X., Ding, Y., Ke, Q.: Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks. Emnlp pp. 3901–3910 (2018), http://aclweb.org/anthology/D18-1424
35. Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., Zhou, M.: Neural question generation from text: A preliminary study. In: National CCF Conference on Natural Language Processing and Chinese Computing. pp. 662–671. Springer (2017)