On the Linguistic and Pedagogical Quality of Automatic Question Generation via Neural Machine Translation

Tim Steuer(✉), Leonard Bongard, Jan Uhlig, and Gianluca Zimmer

Technical University of Darmstadt, Darmstadt, Hesse, Germany
tim.steuer@kom.tu-darmstadt.de

Abstract. Allowing learners to self-assess their knowledge through questions is a well-established method to improve learning. However, many educational texts lack a sufficient amount of questions for self-studying. Hence, learners read texts passively, and learning becomes inefficient. To alleviate the lack of questions, educational technologists investigate the use of automatic question generators. However, the vast majority of automatic question generation systems consider English input texts only. Therefore, we propose a simple yet effective multilingual automatic question generator based on machine-translation techniques. We investigate the linguistic and pedagogical quality of the generated questions in a human evaluation study.

Keywords: Automatic question generation · Self-assessment technologies · Educational technology

1 Introduction

Reading is a crucial way of learning. However, readers often do not learn efficiently with texts, due to possibly misread facts or missing conceptual connections. Thus, reading texts passively is not enough to fully understand the texts' content. To help learners understand even difficult texts, actively engaging them is a useful teaching method. Hence, adjunct questions may improve the learning outcome [1]. However, many texts lack an appropriate amount of questions needed for efficient learning. State-of-the-art natural language processing techniques allow the generation of factual questions on given texts with minimal manual intervention. They receive a sentence or short paragraph and a desired answer. Given the inputs, they transform them into an appropriate question asking about the desired answer. In education, such Automatic Question Generators (AQGs) could alleviate the lack of textbook questions by generating questions on-the-fly, increasing learning efficiency. However, state-of-the-art neural network-based methods mainly work in English. Hence, in this article, we would like to address how to transfer an English AQG into other languages via neural machine translation (NMT). We investigate the linguistic and pedagogical qualities of the generated questions and explore the following research

T. De Laet et al. (Eds.): EC-TEL 2021, LNCS 12884, pp. 289–294, 2021.
https://doi.org/10.1007/978-3-030-86436-1_22

question (RQ): To what extent can a combination of NMT and AQG be utilized to generate linguistically and pedagogically sound questions about texts?

2 Related Work

The research field around AQG has recently shifted from rule-based (e.g. [7]) and template-based approaches (e.g. [10]) towards neural network-based models [5]. The main reason is the superior language generation capabilities of neural models compared to rule-based models [5]. State-of-the-art systems apply large-scale language model pre-training before fine-tuning on SQuAD to improve the general English token prediction capabilities before predicting the task [4]. In education, the LearningQ [2] and RACE [8] datasets are two large-scale datasets comprising questions with explicit educational intent. Furthermore, there have been studies exploring the educational use of neural question generation approaches in empirical studies [6,15,16]. Besides the English language, there also exist AQGs in other languages. For Japanese, using sequence-to-sequence learning and classical statistical learning techniques as AQG have been explored on a small-scale dataset with promising results [12]. Another study uses the over-generate and rank approach in Chinese, combining a rule-based system with a statistical ranking in a factual AQG, outperforming the rule-based system [9]. Recently XNLG [3] has been proposed, a multilingual neural language model pre-trained in fifteen languages. For Chinese, it has been demonstrated that if XNLG is fine-tuned for AQG on SQuAD, it can generate plausible questions [3]. However, to achieve those results, XNLG was pre-trained using computational resources often not available to researchers on a large-scale multilingual corpus [3]. If the corpus does not contain the language target for generation, there is currently no way to import it into XNLG later. Hence, in such cases, pre-training needs to be repeated. Consequently, although the multilingual model may be a good option for some languages, it heavily relies on computational resources and cannot easily be extended with novel languages later.

3 Approach

The proposed approach combines two NMT models and an AQG model. The general idea is to start with a text in a source language, translate it to English, apply the AQG on the translated text, and finally back-translate the resulting question. The input consists of small paragraphs written in the source language, and the output contains the generated question also written in the source language. The AQG takes a paragraph and an answer-candidate as input. The answer-candidate is needed to specify which question we aim to generate. Hence, before generating with the AQG, an answer-candidate must be selected. While the answer-candidate selection is an active field of research and multiple methods exist to detect promising answer-candidates in a text (e.g. [16,17]), we opt for a relatively simple approach for our initial experiments. We extract the longest noun phrase in the paragraph, expecting that it carries the most information

in the small paragraphs. Having the answer-candidate and input sentence, we translate them into English using the WMT19 NMT model by Ng et al. [11]. The NMT model consists of transformer-based neural architecture that is trained in large-scale on filtered parallel corpora. It is built on top of the open-source FAIRSEQ sequence modelling toolkit and is developed to be used in research and production [13]. The neural AQG employed in our work is UNILM by Dong et al. [4]. The UNILM model consists of a transformer-network that is pre-trained on a large corpus using multi-task learning with different sequence masks [4].

4 Empirical Evaluation

The evaluation study comprises two human annotators with a background in educational sciences, annotating 80 generated questions in random order. The questions are generated for 48 educational texts, written by teachers based on Wikipedia articles, as first used by Rüdian et al. [14]. We randomly select 26 texts with a total of 80 questions from the initial 48 articles. The dataset has question-worthy sentences in the texts already marked by educational experts. During data annotation, annotators apply the hierarchical scheme by Horbach et al. [6]. We achieve a average Krippendorff's $\alpha = 0.35$ over the nine categories.

An overview of the results in the levels of the annotations scheme is given in Fig. 1. We report the results relative to the remaining questions and to all questions, since some questions will not be annotated fully, due to the hierarchical nature of the evaluation scheme. Hence the reported metrics have the form in the form *Relative% (All%)*. The provided percentages are averaged over the two annotators and it is important to notice that the bars are relative to the remaining questions in an evaluation level and not to all questions. In total, 88% of the questions are rated as semantically meaningful and understandable. The remaining 12% are rated as not understandable. We found 93% (82%) of the questions to be related to the texts' domains, and 97% (85%) of the questions to be free of language-errors. The most problematic factor on the second level was the clarity of the generated questions with: *clear* 52% (46%), *more or less clear* 32% (28%) and *not clear* 16% (14%). The majority of questions is answerable 80% (59%) with only 20% (15%) being unanswerable. Furthermore 63% (46%) of questions are accepted without rephrasing whereas 37% (27%) should be rephrased. The questions usually ask for information directly given in one position of the text 62% (35%) or in multiple positions in the text 36% (21%). Using additional external knowledge is only rarely needed 2% (1%). The information inquired is central in 82% (48%) and not central in 18% (11%). Finally, 43% (25%) of the questions would be used, 32% (19%) would *maybe* be used and 26% (15%) would not be used in an educational setting according to our experts.

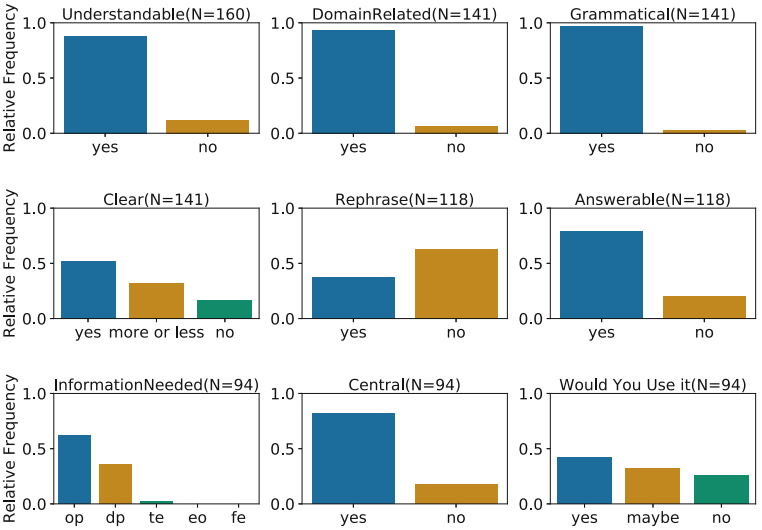

Fig. 1. The results obtained in the evaluation study. The N indicates how many questions where remaining for the bar plot of the given evaluation item.

5 Conclusion

In regard to our research question, the results highlight the important properties of the generated questions. First, the syntactic linguistic quality of the generated questions is high. The majority is understandable and free of language-errors. A look in the data suggests that the incomprehensible questions often stem from a faulty translation of domain terminology. Second, the semantic linguistic quality, the clarity of questions seems to be of concern. Although the questions were mainly answerable and domain-related, annotators found them often to be only more or less clear. In terms of the pedagogical quality, we are able to generate central factual questions aiming directly at the knowledge stated in the text in roughly 50% of all inputs. Our annotators report that they may use around 34% of all generated questions in an educational setting. Consequently, we conclude the approach generated too many pedagogically unimportant questions to give them directly to learners. Currently, we only applied basic answer selection and did not use any question re-ranking method. Related work has shown that answer selection [16] or question ranking [9] may influence the output quality significantly. In future work, we will thus explore these methods. Given sufficient question quality, the approach could then be used to recommend questions to a respective passage to a teacher. The teacher quickly evaluates which of them make sense and marks the good ones as additional self-assessment material with the click of a button.

Acknowledgements. This work is funded by the Hessian State Chancellery in the Department of Digital Strategy and Development in the Förderprogramm Distr@l (Förderprodukt: Digitale Innovations- und Technologieförderung, Förderlinie: 2A Digitale Innovationsprojekte).

References

1. Anderson, R.C., Biddle, W.B.: On asking people questions about what they are reading. In: Psychology of Learning and Motivation, vol. 9, pp. 89–132. Elsevier (1975)
2. Chen, G., Yang, J., Hauff, C., Houben, G.J.: LearningQ: a large-scale dataset for educational question generation. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
3. Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X.L., Huang, H.: Cross-lingual natural language generation via pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 7570–7577 (2020)
4. Dong, L., et al.: Unified language model pre-training for natural language understanding and generation. arXiv preprint arXiv:1905.03197 (2019)
5. Du, X., Shao, J., Cardie, C.: Learning to ask: neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106 (2017)
6. Horbach, A., Aldabe, I., Bexte, M., de Lacalle, O.L., Maritxalar, M.: Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 1753–1762 (2020)
7. Huang, Y., He, L.: Automatic generation of short answer questions for reading comprehension assessment. Nat. Lang. Eng. **22**(3), 457 (2016)
8. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: Race: large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683 (2017)
9. Liu, M., Rus, V., Liu, L.: Automatic Chinese factual question generation. IEEE Trans. Learn. Technol. **10**(2), 194–204 (2016)
10. Mazidi, K., Nielsen, R.D.: Pedagogical evaluation of automatically generated questions. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 294–299. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07221-0_36
11. Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., Edunov, S.: Facebook fair's WMT19 news translation task submission. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 314–319 (2019)
12. Nio, L., Murakami, K.: Intelligence is asking the right question: a study on Japanese question generation. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 771–778. IEEE (2018)
13. Ott, M., et al.: fairseq: a fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038 (2019)
14. Rüdian, S., Heuts, A., Pinkwart, N.: Educational text summarizer: which sentences are worth asking for? DELFI 2020-Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik eV (2020)
15. Steuer, T., Filighera, A., Rensing, C.: Exploring artificial jabbering for automatic text comprehension question generation. In: Alario-Hoyos, C., Rodríguez-Triana, M.J., Scheffel, M., Arnedillo-Sánchez, I., Dennerlein, S.M. (eds.) EC-TEL 2020. LNCS, vol. 12315, pp. 1–14. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57717-9_1

16. Steuer, T., Filighera, A., Rensing, C.: Remember the facts? Investigating answer-aware neural question generation for text comprehension. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 512–523. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_41
17. Willis, A., Davis, G., Ruan, S., Manoharan, L., Landay, J., Brunskill, E.: Key phrase extraction for generating educational question-answer pairs. In: 2019 Proceedings of the Sixth ACM Conference on Learning@ Scale, pp. 1–10 (2019)