# Remember the facts? Investigating Answer-aware Neural Question Generation for Text Comprehension

Tim Steuer (✉)[0000−0002−3141−712X], Anna Filighera, and Christoph Rensing

Technical University of Darmstadt, Germany
{tim.steuer,anna.filighera,christoph.rensing}@kom.tu-darmstadt.de

**Abstract.** Reading is a crucial skill in the 21st century. Thus, scaffolding text comprehension by automatically generated questions may greatly profit learners. Yet, the state-of-the-art methods for automatic question generation, answer-aware neural question generators (NQGs), are rarely seen in the educational domain. Hence, we investigate the quality of questions generated by a novel approach comprising an answer-aware NQG and two novel answer candidate selection strategies based on semantic graph matching. In median, the approach generates clear, answerable and useful factual questions outperforming an answer-unaware NQG on educational datasets as shown by automatic and human evaluation. Furthermore, we analyze the types of questions generated, showing that the question types differ across answer selection strategies yet remain factual.

## 1 Motivation

Reading materials encode a significant amount of our human knowledge, from cooking recipes to textbooks about quantum mechanics. When we are learning, we are often relying on those reading materials as our primary source for knowledge acquisition.

Yet, learning by reading is often challenging and text comprehension depends not only on the reader but also on the text. Even advanced readers occasionally experience difficulties while reading. Texts encompassing jargon, assuming a lot of prior knowledge, or using a specific style of writing challenge even the best of readers. Consequently, providing additional text-specific help might be of great value, not only for novices but also for the intermediate and advanced.

An established reading aid is questioning the readers about the content of the text [1,14]. Depending on the type of questioning, it has different effects. Factual questions direct the attention of learners to specific aspects of the text [1], helping them to remember facts easily. Conversely, comprehension questions require learners to combine different aspects of the text, supporting deeper understanding [1]. That is, to get the most benefit from asking readers, combining different types of questions is important [1,10].

Yet, posing questions is a challenging task even for humans. Authors first need to understand the underlying texts. Next, they have to identify meaningful facts and connections, which are important for the learners' understanding. Finally, they have to state a question in such a way that it actually fosters text comprehension. As a result, having well written, manually authored questions in formal learning settings is expensive, and almost impossible in informal learning settings, where the amount of reading materials is endless.

Automatic question generation is a research field investigating how to create questions without human intervention. It is used in different domains such as dialog systems, question answering or in educational settings. Ideally, to foster text comprehension, an automatic question generator receives the reading material, e.g. a text passage, as input and poses meaningful questions about this text, alleviating the need for expensive human questioning.

However, those systems are far from perfect and posing fluent and meaningful questions from unstructured text is still under active research. The current state-of-the-art systems are answer-aware neural generators (NQGs). It has been shown that such systems generate questions with excellent fluency and acceptable relevancy [7].

They are used in dialog systems and to augment question answering data, but are rarely seen in the education domain. During generation they expect two inputs (see Figure 1). First, they generate questions given a single question-worthy sentence (context sentence) instead of the whole unstructured reading material. Second, they use an explicitly marked expected answer inside the given context sentence (answer candidate).

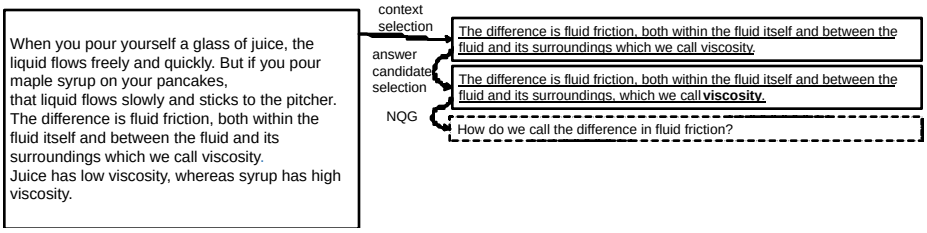


**Fig. 1.** Automatic question generation by selecting the context sentence (underlined) and the answer candidate (bold) from a physics paragraph before generating the actual question via an answer-aware NQG.

This paper makes two contributions. First, we apply answer-aware NQGs to texts in the educational domain and investigate the quality of the generated questions by conducting automatic and human evaluation. Furthermore, we propose two novel answer candidate selection strategies, relying on semantic graph matching, which are easily adaptable to different cases.

## 2 Related Work

The following section will examine the problem of question generation from different viewpoints. It aims to exemplify the challenges of the task and to motivate our design decisions. For a thorough review of automatic question generation in education we refer to Kurdi et al. [17] and for a general review of NQGs to Pan et al. [22].

The literature usually distinguishes three types of automatic question generation approaches. The most common in the field of educational research are rule-based and template-based approaches [17], while outside of the educational domain NQGs are state-of-the-art [22]. Systems in the educational domain investigate a variety of different question types such as Gap-fill questions, multiple-choice questions or Wh-questions in a variety of domains such as generic text comprehension, history or biology [17]. They rely either on text [19] or structured data such as ontologies or knowledge-bases [15] for their context and answer candidate selection. When relying on text, the answer candidate selection of the systems is mostly done via shallow semantic parsing such as semantic role labeling or named entity recognition [12,21]. Furthermore, some authors train classifiers on human-annotated data [16,2].

Looking outside the educational domain, NQGs evolved from relatively simple sequence to sequence models, relying only on the context sentence and the statistical regularities of language to generate questions [9], to sophisticated model with different facets. Subsequent systems make use of advanced neural architectures [7], take desired answers into account [7,25] and are difficulty-aware [13]. These neural approaches have been shown to be superior in terms of naturalness and grammatical correctness by automatic and empirical measures [9,22]. Current state-of-the-art systems are answer-aware NQGs, outperforming answer-unaware and non-NQG approaches [7,22].

Looking at the application of NQG systems in educational settings, relatively little work has been done. Recently, datasets have been collected, containing questions on different cognitive levels, providing more training data for NQGs in education [5,18]. Initial experiments on those datasets have shown that answer-unaware NQGs also outperform rule-based systems on those datasets [5]. Furthermore, selecting the question-worthy context sentences from text either by using classifiers [8] or relying on methods of extractive summarization [4] has been investigated. Preliminary results show that none of the investigated algorithms consistently performs best on all datasets, with LexRank [11] being one of the best performing approaches.

## 3 Research Questions

Our research is guided by the related work and the fact that answer-unaware NQGs outperform rule-based systems on educational datasets and answer-aware NQGs outperform all other systems on non-educational datasets. Thus we hypothesize answer-aware systems will also perform better for educational scenarios, leading to our first research question (RQ1):

1. To what extent are answer-aware NQGs more useful in educational scenarios than answer-unaware NQGs?

Aside of this direct comparison, more nuanced analysis is also important as we need to pose different question types to the learner to achieve optimal support. Therefore, the interaction between the NQG and the answer selection has to be investigated. Only asking for plain facts will not result in the best learning outcome and we hypothesise that some answer selection strategies yield more factual questions than others. Additionally, We assume that answer selection methods have a strong influence on some but not all quality criteria. We suspect that the grammaticality of the question is not altered by using different strategies but that the usefulness of the generated questions and their respective question types (e.g. what vs. why questions) is influenced by different answer selection strategies. We therefore pose our second research question (RQ2):

2. How do different answer selection strategies influence types and quality aspects of the generated questions?

We operationalize RQ1 and RQ2 by looking at the grammaticality, the answerability and the usefulness of the generated questions. Grammaticality is necessary for a question to be comprehensible at all. Furthermore, high grammaticality results in a more fluent reading of the question. We understand answerability as, how well can the answer to the generated question be given taking into account only the context sentence that was used to construct it. This score not only indicates whether the question is meaningful at all, but also whether the answer selection and question generation have worked well together. Finally, we are looking at the usefulness of the generated questions. A useful question is one that covers major concepts or fosters text comprehension whereas a useless question does not help to understand the text any better. Thus, this score informs us about the suitability of the generation process for educational purposes.

## 4 Experiment Setting

To investigate our research questions, we implement a question generation process comprising constant context selection and varying answer candidate selections. We compare an answer-unaware NQG baseline with three different answer-aware NQGs, yielding four different conditions in total. For the context selection in all conditions, we learn from Chen et al. [4] and use LexRank.

### 4.1 Answer-unaware Condition

The answer-unaware NQG [9] is the baseline model from the related work [5]. It consists of a sequence to sequence NQG with attention. It rewrites the context sentence to a question, implicitly selecting an answer inside the sentence. Therefore it is answer-unaware, as it does not explicitly need the answer candidate as an input. We train the system on the SQuAD dataset with the same parameters as given in the authors' paper until we reach a similar performance measured by BLEU-4 [23] on the provided validation set.

### 4.2 Nsubj Condition

We select the subject phrase from context sentences as the answer candidate for the question. We choose this strategy because the subject is frequently correlated with the main protagonist in a sentence. Furthermore, it is a common constituent in many sentences and thus can be selected in most sentences as a plausible answer. Finally, we suspect that asking for the subject of a sentence will yield many factual questions asking for the main protagonists of a sentence or story. In other words, when answering such questions, learners are thinking about the main driving forces of a story.

To implement the strategy, the selected context sentence is dependency-parsed [6] using Stanford CoreNLP 3.9.2 [20], resulting in a semantic graph representing the grammatical relationships of the sentence. Next, we use Semgrex matching to extract relevant information from the graph [3]. This has the advantage that we do not have to write complicated graph traversal code to extract vertices that belong to a grammatical relationship. Instead, a Semgrex pattern describes subgraphs with special properties, that can easily be processed further. We apply pattern matching to all nodes under the sentences subject relation. For sentences containing multiple candidates, we heuristically select the longest, under the assumption that longer inputs are beneficial for the question generator. Note that this approach can easily be extended by changing the Semgrex pattern e.g. by matching adverbial clauses and checking the resulting subgraph to only express consequences.

To generate the actual question, an answer-aware NQG [7] based on a neural transformer [24] is used. It is pre-trained on unidirectional, bidirectional and sequence to sequence prediction tasks. For our task, we use the publicly available fine-tuned question generation model[1] provided by the authors, which is a 24-layer, 1024-hidden states, 16-attention heads 340M parameter model trained on Wikipedia and the BookCorpus and fine-tuned on the SQuAD dataset.

### 4.3 Dobj Condition

We select the direct object phrase from context sentences as the answer candidate for the question by using the same algorithm as in the *Nsubj* condition.

Direct objects are also common parts of sentences, allowing the application of this strategy in most cases. Yet, in contrast to the subject, direct objects are more often targets of actions. Hence, we suspect that asking for direct objects will yield questions having different purposes than in the Nsubj condition. Using direct objects as answer input may e.g. cause the NQG to focus more on the carried out action which might be favourable for understanding. The generation of the question is done with the same answer-aware NQG as in the *Nsubj* condition.

[1] https://github.com/microsoft/unilm

### 4.4 Random condition

We apply basic answer candidate selection by selecting one word from the given sentence at random. The sentence is tokenized [2] and a word is sampled at random. As discussed in the related work section, different neural architectures result in different performing generators. Thus, we include this strategy in the experiments to measure the influence of the different neural architectures independent of their answer-awareness. Observing high-scoring metrics when applying this strategy implies that the answer-aware generator's underlying architecture produces better results detached from the answer candidate. The generation of the question is done with the same answer-aware NQG as in the *Nsubj* condition.

## 5 Results

### 5.1 Datasets

We conduct an automatic and a human evaluation. We focus on texts given by the RACE dataset [18]. It is a publicly available educational dataset, comprising passages and questions generated by human experts for the Chinese English reading exams. It covers different domains in middle to high school difficulty.

Moreover, we also report some automatic evaluation results for the TED-ed part of the LearningQ [5] dataset which also covers a wide variety of topics. This dataset is gathered by crawling the transcripts of TED-ed, an educational video provider, and the corresponding comprehension questions posed by educational experts. Albeit we report such results for comparability, we focus on RACE because of the different nature of video transcripts compared to educational texts.

Note that we filter both datasets before conducting our evaluation. We remove all questions not ending with a question mark (e.g. fill-in-the-gap type of questions), resulting in 1089 paragraphs and 5235 gold questions for the LearningQ dataset and 19,944 paragraphs and 40,439 gold questions for the RACE dataset.

### 5.2 Automatic Evaluation

As a proxy for the grammatical quality of the generated questions, we compute BLEU-4 scores similar to Chen et al. [4]. [3] For that, we compare the generated and the gold-standard questions in the given datasets, by only considering the maximum-scoring questions per passage (see Figure 2).

For the RACE dataset, all answer-aware conditions slightly outperform the answer-unaware generator in terms of the BLEU-4. Yet, the differences are marginal except for the Random condition which performs best. For the LearningQ dataset, the Random condition again performs best, however, closely followed by the Answer-unaware condition.

[2] using Stanford CoreNLP 3.9.2
[3] using https://github.com/tylin/coco-caption

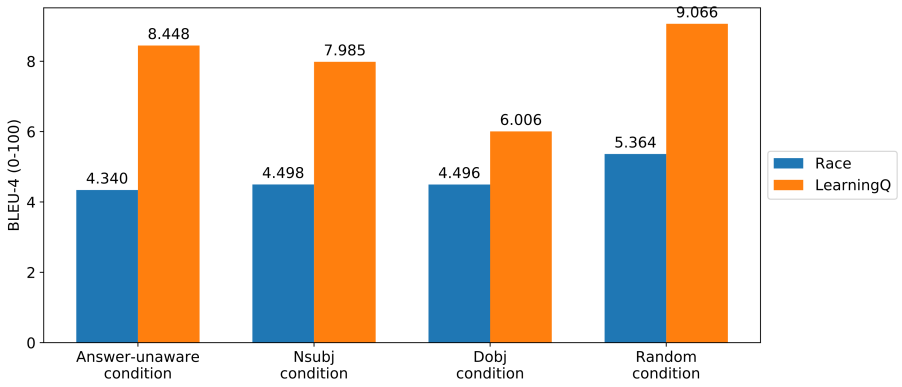


**Fig. 2.** BLEU-4 evaluation results only considering the maximum score per paragraph on the filtered questions of RACE and the TED-ed part of LearningQ.

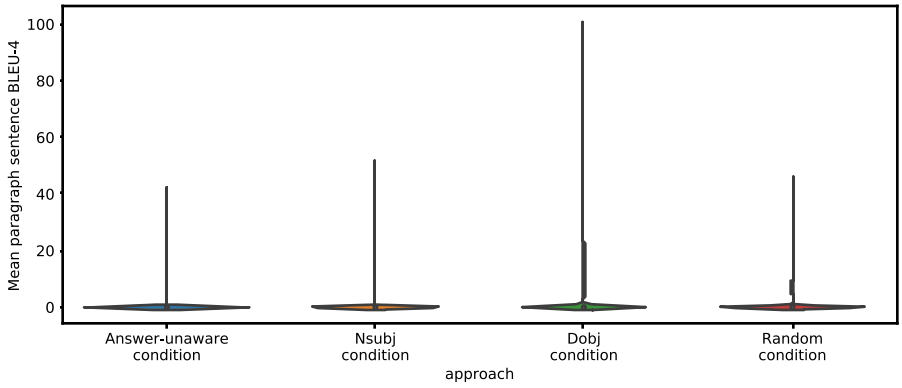


**Fig. 3.** Violin plot of the average sentence BLEU-4 scores per paragraph on the RACE dataset. The estimated kernel density shows that BLEU-4 scores are rarely different from zero.

To validate these results, we compared the distributions of the average sentence BLEU-4 scores per paragraph. They are narrow, with median sentence BLEU-4 scores of zero for any condition (see Figure 3). Put differently, most questions do not overlap with the gold standard. Yet, they nevertheless might be valid as the gold standard comprises only a small subset of all plausible questions. Hence, BLEU-4 may measures little overlap, although the questions are still useful. Second, because the NSubj and Dobj conditions might fail to find an answer-candidate in a context sentence, they generate fewer questions per paragraph than the other two strategies. Only 90% of Nsubj and 65% of Dobj generation attempts succeeded. As a consequence, selecting the maximum scoring sentence per paragraph slightly favors the Random and Answer-unaware conditions, because they always generate the maximum amount of sentences resulting in a higher chance to generate an overlapping question. In summary, although BLEU-4 is often used as a proxy measure for grammaticality, no clear statement about the grammaticality can be made from the automatic measures.

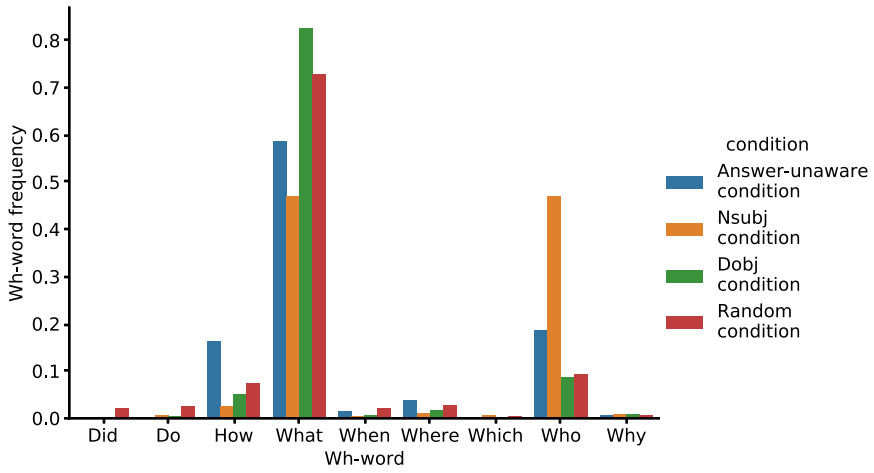


**Fig. 4.** Wh-word frequency on the RACE dataset for the three answer candidate conditions.

When plotting the distribution of the Wh-words we can gather some data for RQ2. The different answer-aware conditions show that they indeed influence the generated question types in our experiments (see Figure 4). The Nsubj condition splits the generated questions almost evenly in "Who" and "What" questions whereas the Dobj and the Random condition mostly pose "What" questions. Hence, looking at these automatically computed statistics provides evidence that most of the generated questions are factual, not asking about reasons or deeper explanations. While this is also true for the Answer-unaware condition, it is worth noting that it stated more "How" questions than any other system.

### 5.3 Human Evaluation

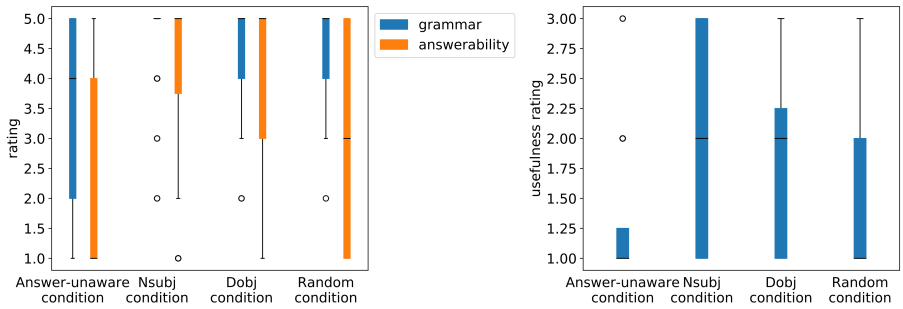


**Fig. 5.** Human evaluation. Left: five-point grammar and answerability ratings. Right: three-point usefulness ratings. The bars indicate the median, the whiskers the 1.5 interquartile range and circles outliers. For Nsubj, the grammar box is a single point.

We conducted a human evaluation with two annotators to get more insights into the grammaticality, answerability, and usefulness of the questions. Both annotators speak English either as a native language or at level CEF[4] B2. We included all four experimental conditions in our evaluation study. We randomly sampled 80 paragraphs from the RACE dataset and assigned each of them to one condition. For every paragraph, we generated three questions. Every annotator evaluated 80 paragraphs having 3 questions each, 240 questions in total. We presented the paragraphs to the annotators in random order. For every paragraph annotators initially saw three context sentences with their generated questions and rated them in terms of answerability and grammaticality on a five-point Likert scale. Then they saw the reading passage together with the questions and rated the questions again for their usefulness (three-point Likert scale) and progressed to the next paragraph. To ensure a common understanding of the scales, an annotation guideline defining answerability, grammaticality, and usefulness was shown. The inter-rater agreement was measured by Krippendorff's $\alpha = .63$ for grammar, $\alpha = .78$ for answerability and $\alpha = .55$ for usefulness. Conflicts were resolved by preferring the native speaker's rating.

The data yields interesting insights into the performance of the different conditions. The median grammaticality rating for the different conditions is 4 in the answer-unaware condition and 5 in all three answer-aware conditions. The median answerability rating is 5 for the Nsubj and Dobj conditions, 3 for the Random condition and 1 for the Answer-unaware condition. The usefulness rating indicates that the Nsubj and Dobj conditions result in a median score of 2 whereas the two other conditions score a median of 1. As shown in Figure 5 most ratings have a non-negligible dispersion.

[4] Common European Framework of Reference for Languages

## 6 Discussion & Future Work

Concerning RQ1, our experiments present evidence that answer-aware NQGs are more suitable for the educational domain than answer-unaware NQGs. The proposed answer selection strategies outperform answer-unaware systems in terms of grammaticality, answerability, and usefulness. For the usefulness criteria, the NSubj and Dobj conditions are the only ones that generate factual questions supporting readers in the median. In contrast, the Answer-unaware condition creates useless questions in the median often even worse than the Random condition. A possible explanation is that the answer-unaware generator mostly selects unimportant information as answers. On the answerability criteria, the answer-aware conditions also perform better, yielding readily answerable questions most of the time. We can rule out the possibility that this is only due to better grammar of the generated questions as the Random condition leads to worse results while scoring high on grammar. For the grammaticality criteria, things are a bit more complex. On the one hand, the BLEU-4 scores are inconclusive. On the other hand, human evaluation shows that the three answer-aware conditions produce more grammatically sound questions. Additionally, the BLEU-4 distributions are almost always close to zero indicating that the gold standard is rarely met. Therefore, we assume that the answer-aware systems are also performing better and that the automatic scores are not representative. However, as the random answer candidate condition also performs quite well on these criteria, answer-unaware systems might profit here from other neural architectures or more training data.

Regarding RQ2, we can see that our strategies result in better questions overall, but the data indicates that the variety of question types is still limited. In every condition, the generated questions remain mainly of factual nature. This is supported by the analysis of the Wh-word distribution, showing that determining questions are posed most often. Furthermore, the usefulness ratings of the annotators indicate that the questions are also mostly of factual nature and not connected to the main ideas of the texts. There might be several reasons for this focus on factual questions. Perhaps the most striking thing is that the whole question generation process currently works on a single sentence basis not taking into account inter-sentence relations. However, important information about the gist of the text can often only be deduced by reasoning about the whole input text. Future work may investigate such reasoning by building NQG processes working with whole paragraphs or extracted summaries, and figuring out synergies between context and answer selection steps. Finally, the used NQGs are mostly trained on data from question answering datasets and thus have most often seen factual questions during their training. In the future, one could explore ways to train or fine-tune such systems on the existing educational datasets.

In summary, this work showed that answer-aware NQGs can generate factual questions to support text comprehension. Yet, more research is needed to pose not only factual but also comprehension questions. Furthermore, we introduced two strategies for answer candidate selection to make the use of answer-aware NQGs possible, which both can easily be extended to more complex patterns.

## 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 (1975). https://doi.org/10.1016/S0079-7421(08)60269-8
2. Blšták, M., Rozinajová, V.: Building an agent for factual question generation task. In: 2018 World symposium on digital intelligence for systems and machines (DISA). pp. 143–150. IEEE (2018)
3. Chambers, N., Cer, D., Grenager, T., Hall, D., Kiddon, C., MacCartney, B., De Marneffe, M.C., Ramage, D., Yeh, E., Manning, C.D.: Learning alignments and leveraging natural logic. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. pp. 165–170. Association for Computational Linguistics (2007)
4. Chen, G., Yang, J., Gasevic, D.: A comparative study on question-worthy sentence selection strategies for educational question generation. In: International Conference on Artificial Intelligence in Education. pp. 59–70. Springer (2019)
5. Chen, G., Yang, J., Hauff, C., Houben, G.J.: Learningq: a large-scale dataset for educational question generation. In: Twelfth International AAAI Conference on Web and Social Media (2018)
6. De Marneffe, M.C., Manning, C.D.: The stanford typed dependencies representation. In: Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation. pp. 1–8 (2008)
7. 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)
8. Du, X., Cardie, C.: Identifying where to focus in reading comprehension for neural question generation. In: EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings (2017). https://doi.org/10.18653/v1/d17-1219
9. Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) (2017). https://doi.org/10.18653/v1/P17-1123
10. Duke, N.K., Pearson, P.D.: Effective Practices for Developing Reading Comprehension. Journal of Education (2009). https://doi.org/10.1177/002205740918900l-208
11. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research **22**, 457–479 (2004)
12. Fattoh, I.E., Aboutabl, A.E., Haggag, M.H.: Semantic question generation using artificial immunity. International Journal of Modern Education and Computer Science **7**(1), 1 (2015)
13. 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)
14. Hamaker, C.: The Effects of Adjunct Questions on Prose Learning. Review of Educational Research (1986). https://doi.org/10.3102/00346543056002212
15. Jouault, C., Seta, K.: Content-dependent question generation for history learning in semantic open learning space. In: International conference on intelligent tutoring systems. pp. 300–305. Springer (2014)

16. Kumar, G., Banchs, R.E., DHaro, L.F.: Revup: Automatic gap-fill question generation from educational texts. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 154–161 (2015)
17. 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)
18. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683 (2017)
19. Liu, M., Calvo, R.A., Rus, V.: Automatic question generation for literature review writing support. In: International conference on intelligent tutoring systems. pp. 45–54. Springer (2010)
20. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010
21. Mazidi, K., Nielsen, R.D.: Pedagogical evaluation of automatically generated questions. In: International conference on intelligent tutoring systems. pp. 294–299. Springer (2014)
22. Pan, L., Lei, W., Chua, T.S., Kan, M.Y.: Recent advances in neural question generation. arXiv preprint arXiv:1905.08949 (2019)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
25. 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