A Comparative Study on Question-worthy Sentence Selection Strategies for Educational Question Generation

Guanliang Chen[1], Jie Yang[2], and Dragan Gasevic[1]

[1] Monash University
[2] University of Fribourg
{guanliang.chen,dragan.gasevic}@monash.edu
[3] jie@exascale.info

Abstract. Automatic question generation, which aims at converting sentences in an article to high-quality questions, is an important task for educational practices. Recent work mainly focuses on designing effective generation architectures based on deep neural networks. However, the first and possibly the foremost step of automatic question generation has largely been ignored, i.e., identifying sentences carrying important information or knowledge that is worth asking questions about. In this work, we (i) propose a total of 9 strategies, which are grounded on heuristic question-asking assumptions, to determine sentences that are question-worthy, and (ii) compare their performance on 4 datasets by using the identified sentences as input for a well-trained question generator. Through extensive experiments, we show that (i) *LexRank*, a stochastic graph-based method for selecting important sentences from articles, gives robust performance across all datasets, (ii) questions collected in educational settings feature a more diverse set of source sentences than those obtained in non-educational settings, and (iii) more research efforts are needed to further improve the design of educational question generation architectures.

Keywords: Educational question generation · Sentence selection · Deep neural network

1 Introduction

In education, automatically generating high-quality questions for learning practices and assessment has long been desired. Previous studies have indicated that reading is one of the most frequent strategies adopted by students to learn [30]. To assess students' understanding, instructors and teachers need to design corresponding assessment questions about the reading material [3, 28]. However, such a question creation process is usually time-consuming and cognitively-demanding. Therefore, *automatic question generation*, which aims at automating the creation process through computational techniques, has attracted much research attention [5, 8, 10, 15].

One common strand of work in automatic question generation is generating questions in a rule-based manner, in which educational experts are recruited to carefully define a set of syntactic rules to turn declarative sentences into interrogative questions [1, 15, 23]. In recent years, deep neural networks have emerged as a more promising approach to question generation [8, 9, 10, 11, 33, 35]. In contrast to rule-based methods, neural question generation methods can capture complex question generation patterns from data without handcrafted rules, thus being much more effective and scalable. Typically, neural question generation methods tackle the generation process as a sequence-to-sequence learning problem, which directly maps a piece of text (usually a sentence) to a question [10].

Table 1. Question-worthy sentence in a paragraph.

Have you ever dropped your swimming goggles in the deepest part of the pool and tried to swim down to get them? It can be frustrating because the water tries to push you back up to the surface as you're swimming downward. **This upward force exerted on objects submerged in fluids is called the buoyant force.**

Given a paragraph or an article, often there are only a limited number of sentences that are worth asking questions about, i.e., those carrying important concepts. An example is shown in Table 1, where the last sentence defines the most important concept "buoyant force". We, therefore, argue that selecting question-worthy sentences is of critical importance to the generation of high-quality educational questions.

Existing studies, however, pay little attention to this step: they either assume that the question-worthy sentences have been identified already [10] or simply take every sentence in an article as input for the question generator. For instance, [15] assumes that all sentences in an article are question-worthy and thus generate one question for each sentence and select high-quality ones based on their linguistic features. To our knowledge, [8] is the only study that explicitly tackles the question-worthy sentence selection problem. It uses a bidirectional LSTM network [19] to simultaneously encode a paragraph and calculate the question-worthiness of a sentence in the paragraph. However, training such a network relies on a large amount of ground-truth labels of question-worthy sentences (e.g., tens of thousands). Obtaining these labels is a long, laborious, and usually costly process. Furthermore, the proposed deep neural network was only validated in short paragraphs instead of the whole article. Considering that reading materials can be much longer and deep neural networks can fail at processing long sequence data due to the vanishing gradient problem [18], it remains an open question whether the proposed method can handle long articles.

Instead of developing a novel neural network architecture that simultaneously does sentence selection and question generation (like [8] does), this work takes one step back and focuses extensively on question-worthy sentence selection. We aim at achieving a better understanding of the effectiveness of different textual features in identifying question-worthy sentences from an article, so as to clar-

ify the main criteria in selecting question-worthy sentences, and to adequately inform question generator design.

To this end, we first propose a total of 9 strategies for question-worthy sentence selection, which cover a wide range of possible question-asking patterns inspired by both low-level and high-level textural features. For instance, we represent our assumption that *informative* sentences are more likely to be asked about by leveraging low-level features such as sentence length and the number of concepts as informativeness metrics; our assumption that *important* sentences are more worth asking about is represented by leveraging semantic relevance between sentences, which can be measured by using summative sentence identification techniques [4, 13]. To evaluate the effectiveness of the proposed strategies, we apply them to identify question-worthy sentences on 4 question generation datasets, i.e., *TriviaQA* [20], *MCTest* [31], *RACE* [21] and *LearningQ* [5]. Among the four considered datasets, only RACE and LearningQ are collected in educational settings and they consist of questions covering various cognitive levels. In contrast, TriviaQA and MCTest are collected to advance the development of machine reading comprehension and they mainly contain questions seeking for factual details. By including all of these datasets, we expect to identify the specific characteristics of question-worthy sentences for the task of educational question generation. We use the sentences identified by the proposed strategies as input for a well-trained question generator and evaluate the effectiveness of sentence selection strategies by comparing the quality of the generated questions.

To the best of our knowledge, our work is the first one that systematically studies question-worthy sentence selection strategies across multiple datasets. Through extensive experiments, we find that *LexRank*, which identifies important sentences by calculating their eigenvector centrality in the graph representation of sentence similarities, gives the most robust performance across different datasets among the nine selection strategies. Furthermore, we demonstrate that questions collected for human learning purposes usually feature a more diverse set of sentences, including those that are most informative, important, or contain the largest amount of novel information, while non-learning questions (e.g., those seeking for factual details from Wikipedia articles in TriviaQA) are often positioned at the start of sentences. Lastly, we show that there is a large improvement space for existing educational question generation architectures.

2 Methodology

Our research methodology is depicted in Figure 1. In the following, we first describe the nine strategies we developed for question-worthy sentence selection and then introduce our method for evaluating these strategies, including the question generator that takes the selected sentences as input for question generation, the experimental datasets, the automatic evaluation metrics, and the human study. We evaluate the effectiveness of the proposed sentence selection strategies by comparing the quality of the questions generated by applying those sentence selection strategies.

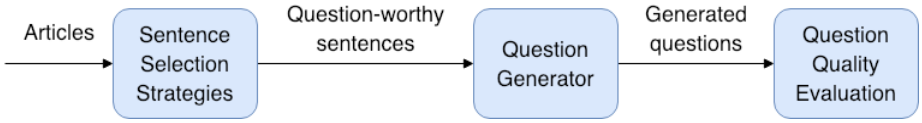

Fig. 1. Research methodology.

2.1 Sentence Selection Strategies

In the following, we describe in detail our proposed sentence selection strategies based on different question-asking assumptions and sentence properties measured by different textual features.

Random sentence (Random). As the baseline, we randomly select a sentence and use it as input for the question generator.

Longest sentence (Longest). This strategy selects the longest sentence in an article. The assumption is that people tend to ask questions about sentences containing a large amount of information, which, intuitively, can be measured by their lengths.

Concept-rich sentence (Concept). In contrast to *Longest*, this strategy assumes that the amount of information can be better measured by the total number of entities in a sentence. The more entities a sentence contains, the richer the information it has.

Concept-type-rich sentence (ConceptType). This strategy is a variant of *Concept*. It calculates the total number of entity types in a sentence to measure the informativeness of a sentence.

The above three strategies approximate question-worthiness of a sentence by *informativeness*, which is further measured by different textual features. In contrast, the following two strategies approximate question-worthiness of a sentence by *difficulty* and *novelty*, respectively.

Most difficult sentence (Hardest). This strategy is built on the assumption that difficult sentences can sometimes bring the most important messages that should be questioned and assessed. Therefore, it chooses the most difficult sentence in an article as the question-worthy sentence. We calculate the Flesch Reading Ease Score [6] of sentences as their difficulty indicators.

Novel sentence (Novel). Unlike *Hardest*, this strategy assumes that sentences with novel information that people do not know before are more question-worthy. We calculate the number of words that never appear in previous sentences as a sentence's novelty score [34] and select the most novel one.

Finally, we introduce three strategies that approximate question-worthiness of a sentence by the relative *importance* of the sentence with respect to the remaining ones in an article. The importance is either measured by the relative position of a sentence or its centrality represented by semantic relevance with other sentences.

Beginning sentence (Beginning). In the research of text summarization, a common hypothesis about sentence positions is the importance of a sentence decreases with its distance from the beginning of the article [26], and therefore less question-worthy. This strategy selects the first sentence in an article as the most question-worthy sentence.

Centroid based important sentence (LexRank). In line with *Beginning*, this strategy also assumes that question-worthy sentences should be selected from those of greater importance. The difference here is that the sentence importance is measured by the centroid-based method, *LexRank* [13], which calculates sentence importance based on eigenvector centrality in a graph of sentence similarities.

Maximum marginal relevance based important sentence (MMR). In contrast to *LexRank*, this strategy computes sentence importance by considering a linear trade-off between relevance and redundancy [4]. That is, the strategy selects the sentence that is most relevant but shares least similarity with the other sentences as the most important sentence.

2.2 Evaluation Method

To evaluate the proposed strategies, we feed the selected sentences to a popular question generator and evaluate the effectiveness on four benchmarking datasets through both automatic evaluation and human evaluation.

Question Generator. There have been several studies working on constructing effective question generators [10, 11, 33, 35], most of which assume that the answer to a question (usually a span of text in the input sentence) is determined prior to the generation of the question and use both the sentence and the answer as input for the question generator. Our focus in this work is to develop effective sentence selection strategies for educational question generation without observing the answers to questions beforehand. We, therefore, adopt the question generator proposed in [10], which only takes a sentence as input to generate a question, as our testbed to evaluate the effectiveness of the proposed sentence selection strategies. The question generator is based on an attention-based sequence-to-sequence learning framework [2], which maps a sentence from an article to a question by using an LSTM encoder and decoder [19]. Particularly, the decoder incorporates the attention mechanism over the encoder hidden states, which enables the question generator to focus on important concepts in the input sentence during question generation.

Datasets. Generally, all datasets with questions and the corresponding articles, which the questions are about, can be used to evaluate the selection strategies. Our work aims at identifying the unique characteristics of sentences that can be used to generate high-quality questions for educational practices, i.e., questions that are natural and readable to people and contain pedagogical value. We, therefore, select experimental datasets that contain natural questions designed by humans instead of search queries [12, 24] or cloze-style questions [16, 17, 25]. We also include datasets that are collected for not only educational purposes

Table 2. Statistics of the 4 experimental datasets.

	TriviaQA	**MCTest**	**RACE**	**LearningQ**
Data source	Wikipedia articles	Human-generated stories	Language learning articles	Lecture transcripts
# Articles	138,537	660	23,530	1,102
# Questions	138,537	2,640	64,120	7,621
# Avg. sentence / article	202.39±173.12	18.93±7.25	19.05±7.05	42.89±25.57
# Avg. words / sentence	27.19±24.92	12.84±6.47	17.36±10.12	19.76±12.48
# Avg. questions / article	1.00±0.00	4.00±0.00	2.73±1.11	6.92±1.80

but also non-educational purposes, e.g., those designed for the advancement of machine reading comprehension, so as to underline the difference between selecting sentences for the generation of learning questions and non-learning questions. To summarize, we include the following datasets for experiments.

- **TriviaQA** [20] contains questions from trivia and quiz-league websites and evidence articles gathered from web search and Wikipedia. Here we only consider questions with evidence articles collected from Wikipedia, which results in 138K question-article pairs.
- **MCTest** [31] consists of 660 stories written by crowd-workers and 2,000 associated questions about the stories.
- **RACE** [21] collects 23,000 reading comprehension articles for English learning and 64,000 assessment questions.
- **LearningQ** [5] contains both instructor-designed questions gathered from TED-Ed and learner-generated questions gathered from Khan Academy. As the learner-generated questions can be redundant about the same knowledge concepts (i.e., same sentences), to avoid concept bias, we only include the 7,000 instructor-designed questions for experiments.

TriviaQA and MCTest are collected in non-educational settings: TriviaQA questions mainly seek for factual details and the answers can be found as a piece of text in the source article from Wikipedia, and MCTest questions are designed for young children. RACE and LearningQ are collected in educational settings: RACE questions are mainly used to assess students' knowledge level of English, whereas LearningQ covers a diverse set of educational topics, more complex articles, and the questions require higher-order cognitive skills to solve. Descriptive statistics of these datasets are given in Table 2.

Automatic Evaluation Metrics. We adopt *Bleu1*, *Bleu2*, *Bleu3*, *Bleu4*, *Meteor* and $Rouge_L$ for evaluation (following [10]). Bleu-n scores rely on the maximum n-grams for counting the co-occurrences between a generated question and a set of reference questions; the average of Bleu is employed as the final score [27]. Meteor computes the similarity between the generated question and the reference questions by taking synonyms, stemming and paraphrases into account

[7]. Rouge$_L$ reports the recall rate of the generated question with respect to the reference questions based on the longest common sub-sequence [22].

Table 3. Examples of generated **Q**uestions and the corresponding source **S**entences.

	Grammaticality	Clarity	Usefulness
S1: As other sources of natural gas decrease, the costs of non-renewable energies rise, and cutting-edge technologies make renewable energies so accessible. **Q1:** *What is the source of natural gas decrease?*	2	1	0
S2: The brain can solve complicated problems, and grasp concepts such as infinity or unicorns. **Q2:** *What problems can the brain solve?*	3	3	0
S3: When testing a new headache medication, a large pool of people with headaches would be randomly divided into two groups, one receiving the medication and another receiving a placebo. **Q3:** *How is a new headache medication tested?*	3	3	1

Human Study. To gain better insights about the quality of the generated questions, we recruited three native speakers of English to rate the quality of 200 randomly-selected questions (along with the corresponding source sentence and article) generated from RACE and LearningQ, respectively. Specifically, we considered three metrics here: *Grammaticality*, *Clarity* and *Usefulness*. Firstly, we only presented a question to the evaluators and asked them to rate the grammatical correctness of the question on a scale of $[1, 3]$, with 3 being exactly correct. Then, the corresponding sentence from which the question was generated was presented to the evaluators and the evaluators were asked to specify how clear the question was and to what extent the question was making sentence given the input sentence on a scale of $[1, 3]$, with 3 being very clear. Lastly, the evaluators were presented with the source article (i.e., the article from which the source sentence was selected) and rate the usefulness of the question for learning (e.g., enabling a better understanding of the article, assessing students' knowledge) on a scale of $[0, 1]$ with 1 being useful and 0 being not useful at all. The ratings given by the three evaluators for each metric were averaged as the final rating for a question. Three examples of the generated questions along with the source sentences and the corresponding human evaluation ratings are given in Table 3. Due to the limited space, we omit the source articles in Table 3.

3 Experiments

This section presents and compares the performance of our proposed strategies for question-worthy sentence selection as evaluated across the considered datasets.

3.1 Experimental Setup

To our knowledge, *SQuAD* [29] is the only dataset which contains ground-truth labels for over 97,000 sentence-question pairs. In line with [10], we also use the labeled input sentences and the corresponding questions in SQuAD for training the question generator. We set the hyper-parameters as suggested in [10] and use beam search ($N = 3$) to generate a question.

Articles can be of different lengths and thus possibly contain different numbers of question-worthy sentences (shown in Table 2). During experiments, the number of selected sentences should be dependent on the number of ground-truth questions gathered about an article: different ground-truth questions are seeking for different details about the article, i.e., based on different question-worthy sentences. We therefore evaluate each of the questions generated by different selected sentences against all the ground-truth questions of a article and consider the result with the best performance as an indication of the selected sentence matched with the ground-truth question.

3.2 Results and Analysis

Table 4 reports the results of the proposed sentence selection strategies on four datasets. We highlight the top-3 strategies for each dataset. Based on these results, several interesting findings are observed as follows.

For TriviaQA, *Beginning* achieves the best performance, indicating that most questions in TriviaQA are about the first sentence in the source article. Considering that the articles of TriviaQA are collected from Wikipedia, such a result can be interpreted by the fact that the first sentences of Wikipedia paragraphs/articles often contain the most important information worth asking about [36]. This observation can be further verified by the well-performing results given by *LexRank* and *MRR* – ranking at the 2nd and 3rd position, respectively – which also identifies important sentences but uses a different method. Overall, these results show that importance-based strategies are more effective than informativeness-based (e.g., *Longest, Concept*), difficulty-based (i.e., *Hardest*), or novelty-based ones (i.e., *Novel*).

For the two datasets collected in educational contexts, namely RACE and LearningQ, *Longest*, *LexRank*, and *Novel* generally show better performance than the other strategies. Such a result suggests that questions in learning related datasets are relevant to a more diverse set of sentences, i.e., those informative, important, or contain novel information, a result is likely due to the diverse learning goals related to the questions. We further observe big gaps between these three strategies and the remaining ones. For example, *Longest*, *LexRank*, and *Novel* are the only strategies achieving *Bleu1* scores greater than 5 and $Meteor_L$ scores greater than 6 on RACE. This observation reveals that sentence selection strategies based on similar sentence properties however measured through different textual features (e.g., *Longest* vs. *Concept* and *LexRank* vs. *Beginning*) can have big variance in terms of performance. This highlights the importance of selecting appropriate textual features in question-worthy sentence selection.

Table 4. Experimental results of automatic evaluation on TriviaQA, MCTest, RACE and LearningQ. The top three results in each metric are in bold.

Datasets	Strategies	Bleu1	Bleu2	Bleu3	Bleu4	Meteor	Rouge$_L$
TriviaQA	Random	6.69	2.07	0.70	0.31	**6.21**	8.29
	Beginning	**9.42**	**3.67**	**1.51**	**0.74**	**6.66**	**10.70**
	Longest	3.37	1.21	0.45	0.21	3.57	8.67
	Hardest	1.99	0.66	0.23	0.11	2.37	6.84
	ConceptMax	0.73	0.19	0.06	0.02	1.63	4.04
	ConceptTypeMax	1.94	0.57	0.19	0.08	2.94	6.02
	LexRank	**8.79**	**3.11**	**1.14**	**0.52**	**5.54**	**9.81**
	MMR	**7.13**	**2.44**	**0.92**	**0.42**	5.06	**8.93**
	Novel	3.25	1.12	0.42	0.20	3.47	8.28
MCTest	Random	4.18	1.41	0.55	0.21	7.04	16.46
	Beginning	4.69	1.56	**0.63**	**0.27**	7.93	**17.36**
	Longest	**5.75**	**1.99**	**0.79**	**0.29**	**9.95**	**17.96**
	Hardest	4.41	1.42	0.51	0.18	7.53	16.73
	ConceptMax	3.92	1.48	0.60	**0.27**	5.72	16.71
	ConceptTypeMax	4.01	1.49	0.60	**0.28**	5.89	16.66
	LexRank	**5.24**	**1.85**	**0.70**	0.22	**8.55**	**18.13**
	MMR	4.53	1.53	0.57	0.22	7.52	17.20
	Novel	**4.92**	**1.58**	0.57	0.20	**9.15**	17.14
RACE	Random	4.24	1.28	0.45	0.20	5.74	11.47
	Beginning	4.50	1.33	0.43	0.18	5.95	11.82
	Longest	**6.48**	**2.08**	**0.74**	**0.33**	**7.83**	**12.84**
	Hardest	4.36	1.35	0.47	0.21	5.51	11.60
	ConceptMax	2.74	0.86	0.34	0.16	3.41	10.69
	ConceptTypeMax	2.78	0.88	0.35	0.17	3.45	10.71
	LexRank	**5.47**	**1.73**	**0.63**	**0.29**	**6.79**	**12.59**
	MMR	4.45	1.39	0.51	0.24	5.78	11.75
	Novel	**5.89**	**1.80**	**0.61**	**0.26**	**7.59**	**12.32**
LearningQ	Random	5.66	1.48	0.43	0.14	5.55	14.83
	Beginning	5.02	1.29	0.37	0.13	5.13	14.53
	Longest	**6.34**	**1.81**	**0.57**	**0.22**	**9.10**	**16.86**
	Hardest	5.92	1.60	**0.52**	**0.21**	5.77	15.48
	ConceptMax	4.57	1.25	0.40	0.16	4.77	14.20
	ConceptTypeMax	4.75	1.29	0.41	0.16	4.91	14.24
	LexRank	**6.74**	**1.91**	**0.62**	**0.26**	**7.44**	**16.40**
	MMR	5.86	1.53	0.47	0.17	5.72	15.12
	Novel	**6.00**	**1.64**	0.50	0.17	**8.93**	**16.28**

Similar results also hold on the MCTest dataset: *Longest*, *LexRank*, and *Novel* generally achieve good performance, which suggests that questions in MCTest are also relevant to a diverse set of sentences. On the other hand, strategies such as *Beginning* and *ConceptMax* also perform well on several metrics, signifying that different measures of sentence properties (e.g., informativeness using *Longest* and *ConceptMax*) do not necessarily lead to highly different sentence se-

lection results on MCTest. Despite this, we can observe that *LexRank* is the only sentence selection strategy consistently ranking in top-3 across all the 4 datasets, demonstrating its robustness against all the other compared strategies.

Table 5. Experimental results of human evaluation on RACE and LearningQ.

	Grammaticality	**Clarity**	**Usefulness**
RACE	2.12	1.92	0.53
LearningQ	1.87	1.34	0.22

Table 5 reports the human evaluation results on 200 questions randomly selected from RACE and LearningQ, whose source sentences were selected by applying the best-performing strategies, i.e., *Longest* and *LexRank*, respectively. Compared to LearningQ questions, RACE questions have higher ratings across all metrics, which indicate that RACE questions are more readable and making better sense to people. This can be explained by the fact that RACE only consists of articles and questions used for English learning, while LearningQ covers a wide range of subjects ranging from arts and humanities to science and technology; as well LearningQ articles are much longer and contain relatively more diverse sentence and question patterns. Correspondingly, this poses more challenges to the question generator to deliver high-quality questions. Noticeably, in terms of *Usefulness*, the ratings are 0.53 and 0.22 on the 0-1 scale for RACE and LearningQ, respectively. This indicates that only about 50% and 20% of the generated RACE and LearningQ questions respectively can be used for educational practices. This is in line with previous findings from [5] and demands further investigations to improve the existing architectures for educational question generation.

4 Conclusion and Future Work

Question-worthy sentence selection is an important however largely ignored topic in automatic generation of educational questions. This paper presented a systematic study on nine sentence selection strategies inspired by different question-asking heuristics. Extensive experiments showed that *LexRank*, which selects important sentences from articles by calculating eigenvector centrality in a graph of sentence similarities, gave robust performance across multiple datasets. Our experimental results also revealed that the beginning sentence in an article is often worth questioning about in non-educational settings, while questions in educational contexts feature a more diverse set of source sentences that are informative, important, or contain novel information. Also, we demonstrated that there is quite some improvement space for developing effective educational question generation architectures. These findings inspire our future research to combine multiple strategies for selecting question-worthy sentences in learning contexts and improve the design of existing educational question generation architectures by applying techniques such as reinforcement learning [32] and generative adversarial networks [14].

Bibliography

[1] Adamson, D., Bhartiya, D., Gujral, B., Kedia, R., Singh, A., Rosé, C.P.: Automatically generating discussion questions. In: AIED (2013)

[2] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

[3] Bahrick, H.P., Bahrick, L.E., Bahrick, A.S., Bahrick, P.E.: Maintenance of foreign language vocabulary and the spacing effect. Psychological Science **4**(5), 316–321 (1993). https://doi.org/10.1111/j.1467-9280.1993.tb00571.x, http://dx.doi.org/10.1111/j.1467-9280.1993.tb00571.x

[4] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. pp. 335–336 (1998)

[5] Chen, G., Yang, J., Hauff, C., Houben, G.J.: Learningq: A large-scale dataset for educational question generation. In: ICWSM (2018)

[6] Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. ITL-International Journal of Applied Linguistics **165**(2), 97–135 (2014)

[7] Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: SMT (2014)

[8] Du, X., Cardie, C.: Identifying where to focus in reading comprehension for neural question generation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2067–2073 (2017)

[9] Du, X., Cardie, C.: Harvesting paragraph-level question-answer pairs from wikipedia. In: Association for Computational Linguistics (ACL) (2018)

[10] Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. In: ACL (2017)

[11] Duan, N., Tang, D., Chen, P., Zhou, M.: Question generation for question answering. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 866–874 (2017)

[12] Dunn, M., Sagun, L., Higgins, M., Guney, V.U., Cirik, V., Cho, K.: Searchqa: A new q&a dataset augmented with context from a search engine. arXiv preprint arXiv:1704.05179 (2017)

[13] Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research **22**, 457–479 (2004)

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)

[15] Heilman, M., Smith, N.A.: Good question! statistical ranking for question generation. In: HLT-NAACL (2010)

[16] Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: NIPS. pp. 1693–1701 (2015)

[17] Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: Reading children's books with explicit memory representations. CoRR **abs/1511.02301** (2015)

[18] Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001)
[19] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
[20] Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In: ACL (July 2017)
[21] Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: Race: Large-scale reading comprehension dataset from examinations. EMNLP (2017)
[22] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL (2004)
[23] Mitkov, R., Ha, L.A.: Computer-aided generation of multiple-choice tests. In: HLT-NAACL (2003)
[24] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268 (2016)
[25] Onishi, T., Wang, H., Bansal, M., Gimpel, K., McAllester, D.A.: Who did what: A large-scale person-centered cloze dataset. In: EMNLP (2016)
[26] Ouyang, Y., Li, W., Lu, Q., Zhang, R.: A study on position information in document summarization. In: COLING. pp. 919–927 (2010)
[27] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
[28] Prince, M.: Does active learning work? a review of the research. Journal of engineering education **93**(3), 223–231 (2004)
[29] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100, 000+ questions for machine comprehension of text. In: EMNLP (2016)
[30] Rayner, K., Foorman, B.R., Perfetti, C.A., Pesetsky, D., Seidenberg, M.S.: How psychological science informs the teaching of reading. Psychological science in the public interest **2**(2), 31–74 (2001)
[31] Richardson, M., Burges, C.J., Renshaw, E.: Mctest: A challenge dataset for the open-domain machine comprehension of text. In: EMNLP. pp. 193–203 (2013)
[32] Sutton, R.S., Barto, A.G.: Introduction to reinforcement learning, vol. 135. MIT press Cambridge (1998)
[33] Tang, D., Duan, N., Qin, T., Yan, Z., Zhou, M.: Question answering and question generation as dual tasks. arXiv preprint arXiv:1706.02027 (2017)
[34] Tsai, F.S., Tang, W., Chan, K.L.: Evaluation of novelty metrics for sentence-level novelty mining. Information Sciences **180**(12), 2359–2374 (2010)
[35] Wang, T., Yuan, X., Trischler, A.: A joint model for question answering and question generation. arXiv preprint arXiv:1706.01450 (2017)
[36] Yang, Y., Yih, W.t., Meek, C.: Wikiqa: A challenge dataset for open-domain question answering. In: EMNLP. pp. 2013–2018 (2015)