Educational Text Summarizer: Which sentences are worth asking for?

Sylvio Rüdian[1], Alexander Heuts[2] and Niels Pinkwart[3]

Abstract: Many question generation approaches focus on the generation process itself, but they work with single sentences as input only. Although the state of the art of question generation's results is quite good, it cannot be used practically as the selection which sentences are worth asking for in an educational setting is currently not possible in an automated way. This limits the ability to generate interactive course materials at scale. In this paper, we conduct a study where we compare teachers' sentence selections of texts with 9 algorithms to find the most appropriate ones concerning reading comprehension. 30 teachers compared the "winner" algorithm, Edmundson with LexRank, which was found to be the optimal algorithm according to previous literature. The result shows that Edmundson outperforms LexRank.

Keywords: Question generation; Online courses; Text summarization.

1 Introduction

Asking questions is one of the main methods used for testing reading comprehension. The preparation of questions is time-consuming and requires an understanding of used texts. The success of interactive online courses depends on a well-structured presentation of contents as well as asking appropriate questions to engage students and test their knowledge.

On the one hand, automatic question generators are available and perform well, but existing approaches of question generators typically focus on generating questions for single sentences based on templates or other methods. However, in many educational scenarios, there is textual material existing that covers many sentences and not just one. The existing question generation techniques cannot select the sentences that are worth asking for in an educational context. This selection task is of critical importance for generating high-quality questions for educational purposes [CYG19]. Doing this at scale can support the process of creating interactive learning material including questions and quizzes.

On the other hand, there are already existing extractive text summarizers that can summarize texts – a question generator could generate questions for selected sentences of the

[1] Humboldt Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Weizenbaum Institute for the Networked Society, ruediasy@informatik.hu-berlin.de, https://orcid.org/0000-0003-3943-4802

[2] Humboldt Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Weizenbaum Institute for the Networked Society, alexander.heuts@hu-berlin.de, https://orcid.org/0000-0002-1755-8970

[3] Humboldt Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, pinkwart@hu-berlin.de

summary. Yet, the resulting summaries depend on the used algorithms, which are designed by using different approaches. Resulting summaries are thus different. Furthermore, the overall goal of text summarizers is to summarize texts by extracting the main aspects. This approach does not necessarily lead to educationally meaningful sources for question generation. To assist instructors and teachers in creating questions for texts concerning reading comprehension, there is the necessity to use an approach that extracts sentences that are worth asking for first and then generate the questions. If an automated summary algorithm produces results that are similar to a teacher's selection, this can thus help to create educationally meaningful questions at scale.

The research objective of this paper is to explore which already existing extractive text summarizers were preferred in sentence selection to ask questions for reading comprehension concerning different subjects that are taught at school.

2 Related Work

Question generation (QG) is classical a sequence-to-sequence problem, with a text or a sentence as input and a question as output. Sometimes, the answer can also be a part of the output. Two approaches emerged in recent years to generate questions: rule-based and methods based on neural networks [De18]. Rule-based methods use templates that could be created manually or be generated. At the generation step, they use the syntactic structure of the language to create questions by re-ordering words. The success of these approaches depends on the existence of multiple templates with diverse structures. All QG approaches are limited to generate questions for single sentences, which still require the selection of sentences that are worth asking for. Using the state of the art, generators create one or more questions for every sentence of a given text, without selecting appropriate ones. Current researchers use human evaluation only to rate generated questions by quality, clarity, usefulness, or grammatical correctness [CYG19, DSC17]. But they limit their evaluations to already selected sentences and mix up the evaluation of selected sentences and the resulting generated questions.

Du et al. [DC17] proposed a hierarchical neural sentence-level sequence tagging-model as a data-driven approach to identify sentences that are question-worthy by using a paragraph as input. For training, they used SQuAD to find sentences that are worth asking for. But SQuAD was designed for "machine comprehension of texts" [Ra16] and not for educational purposes. The main idea of machine comprehension is to ask questions to the machine and the algorithm can find the answer within a given text [Ra16]. It also can be used to train a question generator based on single sentences [HS16]. Concluding to be useful for selecting question-worthy sentences is "not suitable for question generation in the learning context" [Ch18] as it was created by crowdworkers, not by teachers with a pedagogical background and without the aim to create questions for reading comprehension at school. The didactical background for asking questions is missing in the dataset. Assuming that the dataset can be used for real-world educational environments cannot be justified

[Ch18]. “Crowdworkers were tasked with asking […] questions on the content of that paragraph” [Ra16], but a concept to ask questions with the aim of human reading comprehension (e.g. at school) is missing. Although the approach of the authors is a novelty in research, the results need to be revised by using a new database.

G. Chen et al. [CYG19] used four different datasets (TriviaQA [Jo17], MCTest, RACE, and LearningQ) to examine which existing text summarizers select the sentences which were asked for in the samples. All datasets were designed for different purposes. RACE is mostly used in the setting of language learning. LearningQ contains different educational questions of different subjects with 7k questions created by instructors, containing questions to seek for an in-depth understanding of taught concepts. The remaining two were not created for educational settings. The question generator itself was trained with SQuAD [Ra16]. In their evaluation, they used Grammaticality, Clarity, and Usefulness to evaluate generated questions by humans and found out that LexRank performs best.

The overall problem that datasets were used to train models for purposes that they were not created for can result in recommendations that are practically unfeasible. An evaluation of resulting generated questions depends on the quality of the question generator itself, which limits the aimed evaluation of the selection part. In this paper, the main concern is to focus on the selection of sentences for the sake of human reading comprehension to support teachers and instructors to create or generate questions. We focus on two major research questions: 1) Which text summarizer is most similar to the instructor selections for the task of reading comprehension? 2) Does this text summarizer perform better than LexRank, which is the best according to the literature [CYG19]?

3 Methodology

3.1 Study Design

Our study consists of a pre-study, followed by a comparison of our findings with the literature's findings to determine the best state of the art algorithm that selects sentences that are worth asking for in an educational environment. Within the pre-study, we compared selected sentences by teachers and automatic text summarizers. We used 48 texts that have been annotated by 3 teachers each to select sentences that are worth asking for concerning reading comprehension. For this study, teachers were requested by an open call for participation to annotate 8 texts each. For this purpose, German-speaking teachers were sought regardless of their subject or federal state. In sum, 18 teachers annotated all texts, each one three times. Besides, teachers also had to formulate a question for each selected sentence with the corresponding answers so that the task was authentic and they were aware of the reasons for selecting sentences. The texts have an average length of 2687 characters, ranging from 1759 to 3731. They are selections of 12 school subjects (German, geography, history, mathematics, physics, chemistry, biology, computer science, art, music, political education, and theories on sports). For each subject, four specific texts were chosen, that are usually taught at school. The topics of the individual texts are based on the curricula of secondary school subjects from different federal states in Germany. There was no fixed pool of topics from which each subject was chosen randomly. The selection of the topics was made by the instructors, whereby as many different topics as possible were chosen for each subject. Furthermore, the teachers had no influence on the selection or assignment of the topics and texts they were supposed to annotate, as this has been done randomly. Instead of actual school readings, we used extractions of German Wikipedia articles due to copyrights and publishing reasons. The study was planned to be done online, thus using texts by publishing houses was not possible. Texts have been shortened and re-formulated by instructors of the institute for computer science education at the Humboldt University in Berlin, who have partly completed teacher training and are also educating teachers, with the aim to prepare our texts for teaching at school. Strongly interlaced sentences were split into multiple sentences if it was possible. This step was necessary; otherwise, long sentences could be favored as they might contain more information that is worth asking for than shorter ones.

In the main study, we used 9 extractive text summarizers which had the task to select three sentences, more precisely: TextRank [MT04], Edmundson [Ed69], LexRank [ER04], LSA [De90], SumBasic [NV05], Luhn [Lu58], MMR [GC98], KL [HV09], and Longest Sentences. The algorithms had to do their selections on the same texts that we used in our pre-study. According to the length of our texts, asking three questions on a remember level at Bloom's taxonomy [Bl65] is appropriate. The number of selected questions is independent of the used summarizers. Details of all summarizers can be found in the references. We compared the similarity between human selections and the algorithmic ones. Finally, we ordered the similarity scores by value to find the best performing summarizer, that can be used as educational text summarizer to find question-worthy sentences. The question generation step by using single sentences as an input to create a question is not part of this paper as this has already been investigated a lot in detail [CYG19]. With this pre-study, we can detect algorithms that chose sentences most similar to the teachers' selections.

3.2 Algorithmic Evaluation

Different metrics are existing that are broadly used to evaluate generated questions or summarized texts. Metrics like Rouge$_L$ [Li04] can be used to evaluate summarized texts with a reference. This identifies the longest subset of co-occurring sequences of the original sentences concerning generated ones. It uses n-grams recall for evaluation with gold standard sentences as references. Bleu[1-n] uses n-grams' co-occurrences, too, with an additional penalty for overly short sentences [Pa02]. Originally this metric was developed to evaluate a machine-translated text with a human's translation. These sequence-to-sequence problems have in common that texts and summaries or translations can be completely different among each other. By focusing on the micro-level, which sentences are worth asking for, selected sentences from a text will not be changed or transformed. Thus, in our case, using n-gram approaches are overwhelmed as they were designed for transformed texts and we use extractions only. Statistical-based metrics like Rouge$_\mathrm{L}$ or Bleu are necessary if the teacher's selection is not known and the comparison takes place between texts and questions. As we have, due to our study design, the knowledge of selected sentences which are worth asking for, we can use the overlap of selected sentences to compare teachers' selections with selected items by each used summarizer. Using this overlap we define our similarity score ($sim_A \in [0{,}1]$), where $sim_A = 1$ means that selected sentences by teachers and the algorithm A are identical. Let Q be the number of texts, N_Q the number of sentences of text Q, and P the number of selected sentences per text. Let C_L be selected sentences by teachers and C_A selected ones by the algorithm. Each sentence of every text Q is defined by its position $i \in \left[1, N_Q\right]$. $C[i]$ is the selection of the sentence C at position i. Then the similarity for a given algorithm A can be calculated by the following:

$$sim_A = \frac{1}{Q}\sum_{T=1}^{Q}\frac{1}{P}\ \sum_{i=0}^{N_Q}(C_\mathrm{L}[i] * \ C_A[i])$$

$$C[i] = \begin{cases} 1\colon\ if\ sentence\ \text{was}\ selected\ at\ position\ i \\ 0\colon\ else \end{cases}$$

In our study, we set $Q = 48$, and $P = 3$ as we used 48 texts, 3 selected sentences. The algorithm that is mostly part of the top 1 or top 2 ranking will be selected to be the best performing one.

3.3 Human Evaluation

To understand, whether the best performing algorithm can be used practically, we compared the best performing algorithm A with the known summarizer "LexRank" as B, which performs best according to the literature [CYG19]. Teachers got the selections of the two algorithms for each text to decide which selection is the most appropriate one for reading comprehension. The positions of both algorithms A and B were randomized during the study. Additionally, they had to select single sentences of the algorithmic selections that are practically unusable. Each text was rated by 13 teachers to find the most appropriate one. Fig. 1 visualizes the approach. We used "sumy" [Be19] to summarize all texts, limited to extract 3 sentences only. For Edmundson, we used the texts' titles as "bonus words" and a list of 566 stop words as additional parameters. For the MMR summarization, we used an existing implementation[4] and adjusted it to get a summary not limited by the number of characters, but for the number of sentences.

[4] https://github.com/vishnu45/NLP-Extractive-NEWS-summarization-using-MMR/blob/master/mmr_summarizer.py

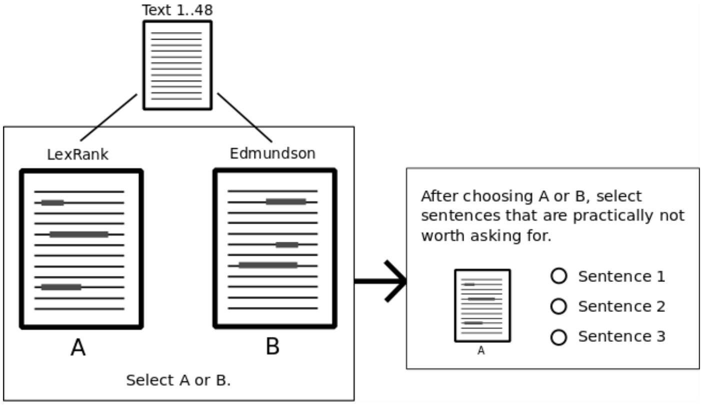
Text 1..48
LexRank
Edmundson
A
B
Select A or B.
After choosing A or B, select sentences that are practically not worth asking for.
A
Sentence 1
Sentence 2
Sentence 3

Fig. 1: Study Design to compare selections of two algorithms.

4 Results

The pre-study explores the similarity of annotations between teachers and the text summarizers. Each text was annotated by three teachers and we reached an inter-rater-reliability score of 0.8 on average. Tab. 1 shows which algorithm creates the most similar selections compared with teacher ones, averaged over all texts as a ranked list. It shows whether the algorithm is the best performing one (Top 1) or the algorithm is at least one of the two most similar ones compared to teacher selections (Top 2). The summarizer "Edmundson" performs best which answers our 1^{st} research question.

Text Summarizer	**in Top 1**	**in Top 2**
Edmundson	0.50	0.61
TextRank	0.11	0.39
LexRank	0	0.33
SumBasic	0	0.22
Luhn	0.11	0.22
LSA	0.06	0.17
Longest Sentences	0	0.33
KL	0.17	0.22
MMR	0.06	0.17

Tab. 1: Text summarizers ranked in pre-study.

According to the literature [CYG19], where LexRank performs best, we cannot confirm this result. In our results, LexRank even not occurs in the Top 1. Thus we aimed to compare the two summarizers LexRank and Edmundson in our main study. We let 30 teachers decide which selections they prefer in a practical setting. Each teacher had to choose the best sentence selection of 16 texts, without telling them that the selections were done by algorithms to avoid prejudice. The same 48 texts of the pre-study were used. We compared the average amount of selections. To find the “winner”, we defined three classes: Edmundson, LexRank, and Ambiguous. The Ambiguous class was chosen if the difference between selections of Edmundson and LexRank was too small. If at least 60% have chosen Edmundson or LexRank, then this was counted to be the preferred version. 37.5% of the selections were ambiguous. By observing the unambiguous selections only, Edmundson was chosen in 73.3% and LexRank was chosen in 26.7%. Thus, Edmundson outperformed LexRank by factor 4.

In addition to the selection of the favored algorithm, all participants were asked to select sentences that they would not use in a practical setting to ask questions about. Teachers selected one of three sentences both for Edmundson and LexRank. This helps to examine the usefulness of using the observed summarizers in real-world scenarios and shows that the summarizer can assist in selecting sentences, but the process still requires an instructor. A complete automatism cannot be achieved at scale yet.

5 Limitations

We limited our study to 48 texts, containing different subjects at school. The base for all texts was Wikipedia and all texts were transformed by instructors to be appropriate for students. To achieve more accurate results, the dataset should be extended to use more texts. Also, texts of publishing houses that are broadly used at schools could be used, but these texts cannot be published and studies cannot be done online due to copyright reasons.

For existing datasets, it is important to publish the purpose, why questions were asked. If the purpose is not known, it is difficult to determine whether a selected sentence is good or not. It is not sufficient that selections were just done by instructors or teachers. Without having access to the specific aim of these questions, it is practically not usable. Other didactical concepts could result in other conclusions, too. Further investigations should consider specifying the purpose as close as possible. Furthermore, we have concentrated on the single sentence selection for reading comprehension, because current question generators can only create questions for single sentences. However, if we want to be even more realistic about reading comprehension, we should be able to ask questions that span sentences for more complex questions [Bl65]. Since this is a much more complex task, we have limited ourselves to single sentences as the first step to this bigger picture.

Text comprehension at school for native speakers is only one concept of using texts. Other methods with different learning goals could be used to extract sentences, that are worth asking for, e.g. in language learning. The resulting optimal algorithm can be different from our findings in other environments. It is important to note that our findings are limited to text comprehension at school. Our result should be compared to other target groups.

6 Discussion

Comparing LexRank and Edmundson, Edmundson uses stigmas words, stop words; title and heading (as “bonus words”); and structural indicators [Ed69]. The algorithm filters uninformative words that contain low semantic information. The main criteria for sentence selection are the frequency of words, cue phrases (that contain words like “significant” or “hardly”), heading words (e.g. the title), and the location of sentences within the text. All these features were used to get a score for each sentence. It was found that resulting summaries overlap with 44% of handcrafted ones. In our study, we also used the title as “bonus words” as suggested by Edmundson. In contrast, LexRank [ER04] is an unsupervised method that finds the most important sentences of a given text by using the concept of sentence salience. It uses a graph-based centrality scoring approach to calculate the importance. All sentences’ similarities can be calculated by the eigenvector centrality.

The success of selecting appropriate sentences by Edmundson depends on the input parameters. If all parameters are empty, the algorithm just selects the first and last sentences of a given text. Thus, it is important to use stop and bonus words. For an automatic approach, it is necessary to have access to the header of a text to be used as bonus word(s). Otherwise, these words have to be selected manually, which limits the practical usefulness. Other approaches (e.g. LexRank) are not dependent on the existence of bonus words. All text summarizers are independent of the text lengths and the amount of the selected sentences. Thus, our approach can be used for smaller and larger texts as well. The instructor can set a fixed number of sentences that should be selected. There is no optimal number, the extractive algorithms create a ranked list of potentially question-worthy sentences and the instructor can be supported with that information. As texts are often shorter in practical settings, our approach can be used for sentence selection.

Although we used German texts and German annotators, the approach can be used for all other languages (e.g. English) that are supported by the observed text summarizers. The results of selected sentences by all summarizers will be the same as long as all parameters are identical. Edmundson also will select identical sentences across different languages as long as used bonus words of the header appear in the same sentences as before. Using synonyms in the header or text can change the result.

As described in the results of our study, we asked all participants to choose their favored selection. Afterward, they were supposed to select sentences which they would not ask for in an educational setting to test the usefulness in practice. On average, one out of three sentences was chosen to be impractical. Although the description was clear - only select sentences that were impractical and select none, if all make sense - most participants selected one out of three sentences, even if all the options were valid. Thus the algorithm could be more optimal than the suggested one, as the majority of our participants selected one sentence to be the worst of three, but not an impractical one.

By asking teachers, which concept they used to select sentences that are worth asking for, concerning reading comprehension in secondary education, only 13.5% said they used a concept in our pre-study. All others have chosen sentences by feeling, what they would ask during classes. This brings us to the conclusion, that they may have selected sentences from the texts more subjectively according to their feelings or that the teachers were working from an implicit understanding of what is important. Experts hold considerable implicit knowledge that they are not explicitly aware they possess. The reasons why teachers have mentioned, that they were not using a concept, can be manifold and need to be further investigated. Teachers that used a concept stated, that they selected sentences that were spread over the text, sentences that contained basic information, definitions, to use the Five Ws [SS11], or to concentrate on different taxonomy levels [Bl65].

Furthermore, it can be discussed, whether there is a bias in the pre-study reasoned at the limitation to exactly 3 sentences since the information content can vary depending on the text. If a text has a very high informational content, it might be difficult for the teachers to be limited to select exactly 3 sentences, because there might be other sentences that have a similar relevance. It is, therefore, possible that there were, for example, 5 relevant sentences in a text, and the teachers also found all 5 important ones but set different priorities in the final decision. If a text had fewer than 3 relevant sentences to which questions could be asked, the teacher would have had to decide on other less relevant sentences by setting the number to 3. Personal priority would also be a factor in this variant. Furthermore, we randomized topics and teachers to get a general view of question-worthy sentences, independent from subjects.

A gap in the training of new teachers could be a reason for missing concepts. If the teachers have not had the opportunity to learn methods and concepts on how to extract important information from texts in a pedagogical context, how to summarize them or how to formulate appropriate questions for text comprehension based on suitable sentences during their training, then these cannot be used by them, as they do not know any.

Another interesting question in this context is whether this gap in qualification between teachers is a general problem or a subject-specific one, since, for example, a teacher of mathematics is comparatively less preoccupied with the text comprehension or extracting the most important information from texts within a pedagogical context than a teacher of language subjects. This does not mean, however, that reading comprehension is not necessary for STEM education, even though this may not be as important as in other subjects. However, even in STEM subjects, longer texts are existing, e.g. on historical or biographical backgrounds. Based on our data, we could only conclude that there was no difference between the teachers' subjects in terms of the use of a concept or method when selecting a sentence. This could be an indicator that neither concepts nor methods were taught to teachers, regardless of the subject.

7 Future Work

A personalized way of finding the sentence extractor which is most similar to the teacher selections can help to find sentences that are worth asking for – based on individual preferences. Therefore, a learning period is required, where annotations of selected questions were collected and compared concerning the summarizer extractions. This methodology can be a first step to generate interactive online course material based on the teaching attributes of individual teachers. This approach can help to create a variety of interactive material that can be different across teachers.

By observing the results, we can say that Edmundson performs better than LexRank in the educational context, but for a practical setting, it is not good enough to be used on a large scale. Edmundson uses semantic information based on the topic and the location of sentences [Ed69]. LexRank is a graph-based centrality scoring but focuses on the frequency and deviation of words, not the content itself. As the best performing algorithm of our study uses semantic features, further investigations should focus on semantic approaches as well.

Currently, no algorithm selects all sentences that teachers will choose. But algorithms like Edmundson can support instructors at finding possibly question-worthy sentences. As we were not able to find the most practical algorithm in 37.5% of all cases, a combination of both text summarizers, Edmundson, and LexRank, could be a solution. Further investigations could observe what the best ratio concerning both algorithms will be to get an optimal result. Additionally, the methodology of Edmundson could be used and adjusted to select sentences that are more appropriate for a learning environment.

Due to a limited budget, we decided to compare LexRank, which was the best performing choice according to the literature with Edmundson, that performed best in our case. A comparison of more pairs of summarizers is required to understand whether these both algorithms are the most appropriate ones for educational purposes. However, if teachers are familiar with methods and concepts for selecting sentences that are worth asking questions about, it is still exciting to investigate why they are not used in practice and in which situations they are more suitable for. Teachers may prefer to follow their feelings and expertise, which they have acquired through experience in the field. Since most of the teachers participating in the study stated that they did not use any concept or specific method in selecting the appropriate sentences, it is still interesting to know what criteria they used to select the sentences otherwise. If we know more about the reasons for the selection of the sentences and why it is worth asking questions about them, suitable algorithms could be adapted even more.

During informal talks after the study, we observed that teachers in their training (in Germany) do not learn how and for which purpose to formulate questions, but rather how to formulate tasks with the help of operators. Operators are formulated as verbs (e.g. describe, represent, explain, compare, ...) and indicate what a student should do in concrete terms. Especially in exam situations teachers should avoid asking questions and instead use operators as keywords because they define specific requirements for students and there can be no misunderstandings as operators are fixed. This can help for further studies to investigate automatic task generation by using operators, instead of questions.

8 Conclusion

In this paper, we explored how text summarizers perform with the task of question-worthy sentence selection in an educational environment for text comprehension at school. Therefore, we compared the sentence selections of teachers to the summarizers' ones. The observation of the overlap with the algorithmic choices has shown that the text summarizer "Edmundson" performed best in our setting. We compared this summarizer to LexRank, which was known to be the best performing one in educational settings [CYG19]. The direct comparison of the summarizers' selections has shown that Edmundson was favored by teachers in contrast to LexRank. To examine the usefulness, we observed which selections are not worth to ask questions about. Although the results of applying extractive summarizers to select sentences that are worth asking for in an educational setting are quite promising, one of three selections was marked to be not useful. This shows that a subset of the algorithmic selections can be useful. The preselection can be helpful to support teachers during the selection process and to reduce the time required. From the practical point of view, a fully automatic selection at scale still requires an instructor.

Acknowledgments: This work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DII116 (Weizenbaum-Institute). The responsibility for the content of this publication remains with the authors.

Bibliography

[Be19] Belica, M.: Automatic text summarizer: sumy 0.8.1, 19 05 2019. [Online]. Available: https://pypi.org/project/sumy/. [Accessed 12 12 2018].

[Bl65] Bloom, B.: Bloom's Taxonomy of educational objectives, in Vol. 1: Cognitive domain, New York, 1965.

[Ch18] Chen, G. et al.: LearningQ: A Large-scale Dataset for Educational Question Generation, in AAAI, 2018.

[CYG19] Chen, G.; Yang, J.; Gasevic, D.: A Comparative Study on Question-Worthy Sentence Selection Strategies for Educational Question Generation, in AIED, Chicago, 2019, pp. 59-70.

[DC17] Du, X.; Cardie, C.: Identifying Where to Focus in Reading Comprehension for Neural Question Generation, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, Association for Computational Linguistics, 2017, p. 2067–2073.

[De18] Devlin, J. et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

[De90] Deerwester, S. C. et al.: Indexing by latent semantic analysis, in JASIS 41, 6, 1990, p. 391–407.

[DSC17] Du, X.; Shao, J.; Cardie, C.: Learning to Ask: Neural Question Generation for Reading Comprehension, in ACL, 2017.

[Ed69] Edmundson, H. P.: New methods in automatic extracting, in Journal of the Association for Computing Machinery, Vol. 16, No. 2,, 1969, pp. 264-284.

[ER04] Erkan, G.; Radev, D. R.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, in Journal of Artificial Intelligence Research 22, AI Access Foundation, 2004, pp. 457-479.

[GC98] Goldstein, J.; Carbonell, J.: Summarization: (1) Using MMR for Diversity-based ReRanking and (2) Evaluating Summaries, in Tipster Text Program Phase II'I: Proceedings of a workshop, Baltimore, Maryland, USA, Association for Computational Linguistics, 1998, pp. 181-195.

[HV09] Haghighi, A.; Vanderwende, L.: Exploring Content Models for Multi-Document Summarization, in Annual Conference of the North American Chapter of the ACL, Colorado, Association for Computational Linguistics, 2009, pp. 362-370.

[HS16] Heilman, M.; Smith, N. A.: Question Generation via Overgenerating Transformations and Ranking, Pittsburgh, Language Technologies Institute, 2009.

[Jo17] Joshi, M. et al.: TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, in Computation and Language, CoRR, 2017.

[Li04] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, in Text Summarization Branches Out, Association for Computational Linguistics, 2004, p. 74–81.

[Lu58] Luhn, H. P.: The Automatic Creation of Literature Abstracts, in IRE National Convention, New York, IBM Journal of Research and Development, 1958, pp. 159-165.

[MT04] Mihalcea, R.; Tarau, P.: TextRank: Bringing Order to Texts, in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Spain, Association for Computational Linguistics, 2004, p. 404–411.

[NV05] Nenkova, A.; Vanderwende, L.: The Impact of Frequency on Summarization, 2005.

[Pa02] Papineni, K. et al.: BLEU: A method for automatic evaluation of machine translation, in Association for Computational Linguistics, Philadelphia, Pennsylvania, ACM, 2002, pp. 311-318.

[Ra16] Rajpurkar, P. et al.: SQuAD: 100,000+ Questions for Machine Comprehension of Text, in Conference on Empirical Methods in Natural Language Processing, 2016.

[SS11] Salgado, S.; Strömbäck, J.: Interpretive journalism: A review of concepts, operationalizations and key findings, Sage Journals, 2011.