Question-Worthy Sentence Selection for Question Generation

Sedigheh Mahdavi[1], Aijun An[1], Heidar Davoudi[2], Marjan Delpisheh[1], and Emad Gohari[3]

[1] York University Toronto, Canada
{smahdavi,aan,mdelpishe}@eecs.yorku.ca
[2] Ontario Tech University, Oshawa, Canada
heidar.davoudi@uoit.ca
[3] iNAGO Inc.,Toronto, Canada
emadg@inago.com

Abstract. The problem of automatic question generation from text is of increasing importance due to many useful applications. While deep neural networks achieved success in generating questions from text paragraphs, they mainly focused on a whole paragraph in generating questions, assuming all sentences are question-worthy sentences. However, a text paragraph often contains only a few important sentences that are worthy of asking questions. To that end, we present a feature-based sentence selection method for identifying question-worthy sentences. Such sentences are then used by a sequence-to-sequence (i.e., *seq2seq*) model to generate questions. Our experiments show that these features significantly improves the question generated by *seq2seq* models.

Keywords: Question Generation(QG) · Sentence Selection

1 Introduction

In recent years, automatic question generation (QG) has attracted a considerable attention in both machine reading comprehension [6, 34] and educational settings [5, 33]. Automatic question generation aims to generate natural questions from a given text passage (e.g., a sentence, a paragraph). There are two main categories of QG methods: *rule-based* approaches [18, 17] and *deep neural network* approaches based on sequence-to-sequence (*seq2seq*) models [6, 37, 29, 36]. Rule-based approaches mainly use rigid heuristic rules to transform the source sentence into the corresponding question. However, rule-based methods heavily depend on hand-crafted templates or linguistic rules. Therefore, these methods are not able to capture the diversity of human-generated questions [35], and also may not be transformed to other domains [33]. Recently, *seq2seq* neural network models [6, 37, 29, 36] have shown good performance to generate better-quality questions when a huge amount of labeled data is available. Moreover, it has been shown that utilizing the paragraph-level context can improve the performance of seq2seq models in the question generation task [6, 36].

This section provides a brief overview about some of the important features that may or may not be on your specific vehicle. for more detailed information, refer to each of the features which can be found later in this owner's manual. *The remote keyless entry (rke) transmitter is used to remotely lock and unlock the doors from up to 60m (197ft) away from the vehicle. press to unlock the driver door. press unlock-symbol again within three seconds to unlock all remaining doors. press to lock all doors. lock and unlock feedback can be personalized*. see vehicle personalization.

Fig. 1. Sample paragraph from car manuals. Green sentences are question-worthy.

Most existing *seq2seq* methods generate questions by considering all sentences in a paragraph as question-worthy sentences [6, 37, 29, 36]. However, not all the sentences in a text passage (a paragraph or an article) contain important concepts or relevant information, making them suitable for generating useful questions. For example, in Figure 1 only the underlined sentences in a sample paragraph from a car manual dataset (one of datasets used to evaluate the proposed method) are question-worthy (i.e., human may ask questions about them), and other sentences are irrelevant. Therefore, extracting question-worthy sentences from a text passage is a crucial step in question generation for generating high-quality questions.

Sentence selection has been investigated for the purpose of text summarization [26, 9, 11], where sentences in a document are ranked based on sentence-level and/or contextual features. However, few works exist for sentence selection for the task of question generation (QG). Recently, question-worthy sentence selection strategies using different textual features were compared for educational question generation [4]. However, these strategies identify question-worthy sentences by considering features individually, which may not be powerful enough to distinguish between irrelevant and question-worthy sentences.

In this paper, we use two types of features: *context-based* and *sentence-based* features to identify question-worthy sentences for the QG task. Given a passage (e.g., a paragraph), our goal is to investigate the effectiveness of using these features for extracting question-worthy sentences from the passage on the QG performance. In addition, we consider using only the question-worthy sentences in a passage as the context for question generation instead of using the whole passage. We incorporate the context into a *seq2seq* question generation model with a 2-layer *attention* mechanism. We conduct comprehensive experiments on two datasets: Car Manuals and SQuAD [24] and show that the proposed question-worthy sentence selection method significantly improves the performance of the current state-of-the-art QG approaches in terms of different criteria.

2 Related work

2.1 Question Generation

Question Generation (QG) can be classified into two categories. (1) rule-based approach [21, 12, 19] and (2) neural network approach [6, 37, 29]. Rule-based

methods rely on human-designed transformation or template-based approaches that may not be transferable to other domains. Alternatively, end-to-end trainable neural networks are applied to the QG task to address the problem of designing hand-crafted rules, which is hard and time-consuming. Du et al. [6] utilized a sequence-to-sequence neural model based on the attention mechanism [1] for the QG task and achieved better results in contrast to the rule-based approach [12] . Zhou et al. [37] further modified the attention-based model by augmenting each input word vector with the answer position-aware encoding, and lexical features such as part-of-speech and named-entity recognition tag information. They also employed a copy mechanism [10], which enables the network to copy words from the input passage and produce better questions. Both works take an answer as the input sentence and generate the question from the sentence accordingly.

Yuan et al. [34] introduced a recurrent neural model that considers the paragraph-level context of the answer sentence in the QG task. Sun et al. [29] additionally improved the performance of the pointer-generator network [27] modified by features proposed in [37]. Based on the answer position in the paragraph, a question word distribution is generated which helps to model the question words. Furthermore, they argued that context words closer to the answer are more relevant and accurate to be copied and therefore deserve more attention. They modified the attention distribution by incorporating trainable positional word embedding of each word in the sentence w.r.t its relative distance to the answer. Zhao et al. [36] improved the QG by utilizing paragraph-level information with a gated self-attention encoder. However, these methods commonly use the whole paragraph as the context. Our method uses only question-worthy sentences in a paragraph as the context.

2.2 Feature and Graph-based Sentence Ranking and Selection

A variety of rich features have been used to score sentences in a text passage for summarization purposes [26, 9, 11, 15]. In [26], the authors summarized these features in two general categories: importance features and sentence relation features. Importance features (e.g, length of a sentence, average term frequency (*Tf–idf*) for words in a sentence, average word embedding of words in the sentence, average document frequency, position of a sentence, and Stop words ratio of a sentence) are considered to measure importance of a sentence individually. Sentence relation features determine the content overlap between two sentences.

In [9], the number of named entities in a sentence was considered as one of sentence importance features. In [23], three types of features: statistical, linguistic, and cohesion, were applied to score sentences for selecting important sentences. Statistical features assign weights to a sentence according to several features: keyword feature, sentence position, term frequency, the length of the word, and parts of speech tag. Linguistic features: noun and pronouns give higher chances for sentences with more nouns and pronouns to be include in the summary. Cohesion features consider two kinds of features: grammatical and lexical. In order to score and extract sentences that best describe the paragraphs, a

graph-based model, TextRank [20] is used. In this approach, a graph is formed by representing sentences as nodes and the similarity scores between them as vertices. By using the PageRank algorithm [3], nodes with higher scores are chosen as the significant sentences of a given paragraph. Another popular method for deriving useful sentences is LexRank [7], which is a graph-based method capturing the sentences of great importance based on the eigenvector centrality of their corresponding nodes in the graph. SumBasic [31] is another algorithm in which the frequency of words occurring across documents determines sentence significance.

To select sentences for question generation, in [4], different textual features, such as sentence length, sentence position, the total number of entity types, the total number of entities, hardness, novelty, and LexRank measure [7] are individually used to extract question-worthy sentences for a comparison purpose. Here, we train a sentense selection classifier by using multiple features including both context-based and sentence-based features.

3 Methodology

Given a text passage (e.g., a paragraph, a section or an article), our task is to select question-worthy sentences from the passage that capture the main theme of the passage, and use the selected sentences to generate questions. In this section, we first introduce a question-worthy sentence extraction method that extracts all question-worthy sentences from a paragraph. Then, we describe how the question-worthy sentences of a paragraph are incorporated into a *seq2seq* model that uses an attention strategy to generate questions. Figure 2 shows the general view of the proposed method.

3.1 Feature-based question-worthy sentence extraction

Inspired by text summarization methods that extract rich features from a text passage (a paragraph or an article) for identifying summary-worthy sentences [26, 9, 11, 15], we develop a new question-worthy sentence selection method. We consider question-worthy sentence selection as a classification task that evaluates each sentence in the passage utilizing context-based and sentence-based features of the sentence.

Given a training data set that contains a set of passages where each passage consists of a sequence of sentences and each sentence is labelled as question-worthy or not, our task is to learn a classifier from the training data that predicts the question-worthiness of a sentence in a passage. To learn such a classifier, we first extract features of sentences in the training data and then train a classifier based on the extracted features. The training data are represented as $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where x_i, y_i, n are the feature vector of the sentence i, its label, and the number of sentences in D, respectively. The classifier finds a mapping function $F : X \rightarrow Y$, where X is the domain of input sentences

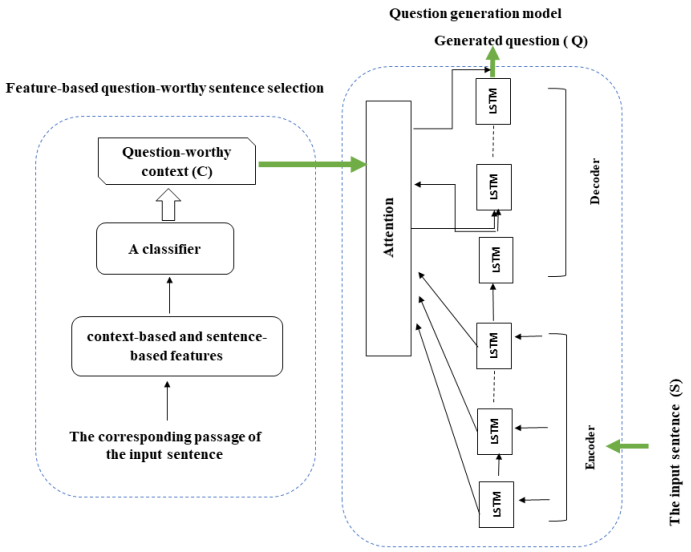

Fig. 2. Proposed framework for question generation

and Y is the set of labels or classes (i.e., question-worthy or not). In our experiment, a Random Forest classifier [13, 2] is trained to identify question-worthy sentences due to its solid performance in text classification tasks, although other classification methods can be used.

We use two groups of features to represent a sentence: context-based and sentence-based features. Context-based features consider the passage which the sequence is in and contain rank features and the *tf–idf* feature. The rank features of a sequence are the ranks of the sentence in its passage obtained from different text summarization methods. The intuition of using rank features is that sentences with important and valuable information contents are ranked higher. Therefore, high rank sentences are more suitable to ask question about. We employ four text summarization methods: TextRank [20], SumBasic [31], LexRank [7], and Reduction [8]. We use four different ranking methods because different ranking methods consider different sets of factors in sentence ranking and all these factors can be considered when incorporating all of them in our sentence representation. The sentence ranks generated by these summarization methods are used as four rank features. To compute the *tf–idf* feature of a sequence, we first compute the *tf-idf* value of each word in the sequence in the context of the passage the sentence is in. That is, the term frequence of a word is the frequency of the word in the sentence and the *inverted document frequency* of the word is the number of sentences containing the word in the passage. We then use the average *tf–idf* value of the words in a sentence as the *tf–idf* feature of the sequence. Intuitively, the *tf-idf* value of a sentence measures the importance of a sentence in its passage.

We also use sentence-based features, which consider only the sentence without its context. Sentence-based features are of two different types: POS-tag (Parts of speech tag) and sentence importance features. Part-of-speech tagging is a basic NLP task that classifies words into their parts of speech and labeling them accordingly. We use six POS-tag features: (1) Number of verbs in a sentence, (2) Number of nouns in a sentence, (3) Number of adjectives in a sentence, (4) Number of adverbs in a sentence (5) Number of pronouns in a sentence, and (6) Number of connection words in a sentence. Our sentence importance features are the length of a sentence and the stop words ratio in a sentence [26].

3.2 Context-aware question generation

We use a *seq2seq* model to generate questions from question-worthy sentences given a passage. In a *seq2seq* question generation model, the objective is to generate a question Q for a text sequence S (e.g., a sentence that answers the question). More formally, the main objective is to learn a model with parameter θ^* given a set of S and Q pairs by solving the following:

$$\theta^* = \arg\max_{\theta} \sum_{Q,S} \log P(Q|S;\theta), \tag{1}$$

Here, we also consider the context of the input sentence S when generating a question from S. We use the question-worthy sentences in the paragraph of sentence S as the context C of S. Thus, our problem is to learn a model with parameter θ^* given a set of tuples $\langle S, C, Q\rangle$, such that:

$$\theta^* = \arg\max_{\theta} \sum_{Q,S,C} \log P(Q|S,C;\theta), \tag{2}$$

To incorporate contexts into the *seq2seq* model, we use the same strategy proposed in [25] for context-aware query reformulation, where a new attention strategy (two-layer attentions) was introduced for incorporating the context of a query into a *seq2seq* model. The model proposed in [25] is called *Pair Sequences to Sequence* (*Pair S2S*) due to the fact that two input sequences are used to generate one output sequence. In the encoder stage of *Pair S2S* model, both the input sequence $S = \{w_t^S\}_{t=1}^M$ and its context $C = \{w_t^C\}_{t=1}^N$ (where w_t^S and w_t^C represent the tth word in S and C, respectively, and M and N are the number of words in S and C, respectively) are separately encoded as follows:

$$u_t^S = RNN^S(u_{t-1}^S, e_t^S) \tag{3}$$

$$u_t^C = RNN^C(u_{t-1}^C, e_t^C) \tag{4}$$

where e_t^S and e_t^C are the word embeddings for the context and the input sentence, respectively. In the decoder stage, the traditional attention mechanism is separately applied on the context and input sequence as follows:

$$c_t^C = \sum_{k=1}^{N} \alpha_{t,k}^C u_k^C \qquad c_t^S = \sum_{k=1}^{M} \alpha_{t,k}^S u_k^S \tag{5}$$

$$\alpha^C_{t,k} = \frac{e^{f(s_t, u^C_k)}}{\sum_{k_i} e^{f(s_t, u_{k_i}{}^C)}} \qquad \alpha^S_{t,k} = \frac{e^{f(s_t, u^S_k)}}{\sum_{k_i} e^{f(s_t, u_{k_i}{}^S)}} \tag{6}$$

where s_t, c^C_t, c^S_t, $\alpha^C_{t,k}$, $\alpha^S_{t,k}$, and f are represents the internal state of recurrent neural network(RNN) at time t, the attention vector for the context, the attention vector for the input sentence, the attention strength for the context, the attention strength for the input sentence, and the attention function, respectively. Then, another attention layer is applied to combine the attention vectors of the input sequence and the context:

$$c^{C+S}_t = \beta_C c^C_t + \beta_S c^S_t \tag{7}$$

$$\beta_C = \frac{e^{f(s_t, c^C_t)}}{e^{f(s_t, c^S_t)} + e^{f(s_t, c^C_t)}} \tag{8}$$

$$\beta_S = \frac{e^{f(s_t, c^S_t)}}{e^{f(s_t, c^S_t)} + e^{f(s_t, c^C_t)}} \tag{9}$$

We apply the above two-layer attentions in [25]. For each input sentence, question-worthy sentences extracted by the feature-based sentence selection method from its corresponding paragraph are considered as the question-worthy context.

Table 1. Evaluation results for important sentence selection on SQuAD. The best results is highlighted in **boldface**.

Method (SQuAD)	Precision	Recall	Accuracy	Macro-F1	Micro-F1
ConceptTypeMax	0.6021	0.3827	0.4679	0.4680	0.4682
ConceptMax	0.6021	0.3827	0.4679	0.4678	0.4681
LexRank	0.7610	0.4836	0.5915	0.5913	0.5916
Emb	0.7000	0.0002	0.3885	0.2801	0.3887
Longest	0.7235	0.4600	0.5620	0.5622	0.5624
FS-SM-IM	**0.8273**	0.6813	0.6405	0.5623	0.6407
FS-SM-Pos	0.6938	0.6920	0.6047	0.5695	0.6049
FS-SM-Rank	0.7283	0.7287	0.6510	0.6206	0.6513
FS-SM	0.7626	**0.7606**	**0.6932**	**0.6658**	**0.6932**

4 Experimental Setup and Results

4.1 Dataset and Implementation Details

We conduct our experiments on the following datasets.

– Car Manual dataset: This dataset (provided by iNAGO Inc. [4]) consists of 4672 QAs created by human annotators from two car manuals (Ford and

[4] http://www.inago.com/

GM). We randomly divided 80% of the dataset into training, 10% validation and 10% test. In this dataset, sentences can be divided into two different classes with label '0' and '1'. Label '1' for a sentence means that humans identify it as a worthy sentence . Sentences with label '0' are irrelevant sentences.

– Processed SQuAD dataset: We use the Stanford Question Answering Dataset (SQuAD) [24], a machine reading comprehension dataset, which offers a large number of questions and their answers extracted from Wikipedia through crowdsourcing. Each example consists of a sentence from an article with its associated question generated by human and its corresponding paragraph. We use this dataset with the same setting as (Du et al., 2017). The data has been split into training set (70,484 question-answer pairs), dev set (10,570 question-answer pairs) and test set (11,877 question-answer pairs).

We train our models with stochastic gradient descent using OpenNMT-py [14], an open source neural machine translation system, with the same hyperparameters as in [6]. The learning rate starts at 1 and is halved at $8th$ epoch. We train a two-layer LSTMs with hidden unit size 600 for 15 epochs.

Table 2. Evaluation results for sentence selection on Car Manuals dataset. The best results are highlighted in **boldface**.

Method (Car manuals)	Precision	Recall	Accuracy	Macro-F1	Micro-F1
ConceptTypeMax	0.6689	0.3679	0.4740	0.4744	0.4747
ConceptMax	0.6690	0.3680	0.4746	0.4747	0.4750
LexRank	0.7508	0.4129	0.5318	0.5328	0.5330
Emb	0.39	0.0002	0.3548	0.2619	0.3548
Longest	0.5436	0.2990	0.3850	0.3855	0.3858
FS-SM-IM	0.5511	0.5706	0.5805	0.5798	0.5808
FS-SM-Pos	0.6576	0.6531	0.6569	0.6572	0.6574
FS-SM-Rank	0.6094	0.6077	0.6189	0.6196	0.6200
FS-SM	**0.7641**	**0.6896**	**0.7150**	**0.7152**	**0.7155**

4.2 Evaluation Metrics

To evaluate sentence selection methods, we use precision, recall, accuracy, and F1 scores. For question generation, we report BLEU-1, BLEU-2, BLEU-3, BLEU-4 [22] and ROUGE-L [16] scores based on the package in [28] for evaluating natural language generation. BLEU-n is a modified precision of n-grams between the reference and generated sentences, while ROUGE-L compares the longest matching sequence of words between system-generated and reference counterparts.

4.3 Question-worthy context Results

We compare our feature-based question-worthy sentence extraction method (FS-SM) with a number of baselines, including LexRank, ConceptTypeMax, ConceptMax, and Longes proposed in [4]. In [4], it was shown that LexRank is

the best question-worthy sentence identification strategy on most datasets. This strategy is based on summary scores of the LexRank [7] summarization method. The ConceptMax and ConceptTypeMax strategies consider the total number of entities and the total number of entity types in a sentence, respectively. In addition, we examine the embedding feature (Emb method) proposed in [26] which represents the sentence content. To analyze the effect of each type of features, we evaluate three variants of FS-SM:

- FS-SM-Pos: A version of FS-SM whose classifier is trained by considering just the POS-tag features
- FS-SM-IM: A version of FS-SM whose classifier is trained by considering just the sentence importance features
- FS-SM-Rank: A version of FS-SM whose classifier is trained by considering just the rank features

Tables 1 and 2 show results on the Car Manuals and SQuAD datasets. The results show that the FS-SM method significantly outperforms the other baselines in terms of classification evaluation metrics. From Tables 1 and 2, it can be seen that all versions of the FS-SM method achieved better results than other strategies.

4.4 Question Generation Results

We compare FS-SM-seq2seq (our QG method) with some baselines for question generation. Tables 3 and 4 show the results for the following QG methods:

- Vanilla seq2seq: The basic *seq2seq* model [30] whose input is a sentence.
- Transformer: Transformer model is a neural network based *seq2seq* model based on the attention mechanism [1] and positional encoding [32]. Its input is a sentence.
- Para-seq2seq: A *seq2seq* model with the 2-layer attention strategy [25] where for each input sentence its whole paragraph is used as its context.
- ConceptMax-seq2seq: A *seq2seq* model with the 2-layer attention strategy [25] that uses the question-worthy sentences identified by ConceptMax from the paragraph of the input sentence as the question-worthy context.
- LexRank-seq2seq: A *seq2seq* model with the 2-layer attention strategy [25] that uses the question-worthy sentences identified by LexRank from the paragraph of the input sentence as the question-worthy context.
- FS-SM-seq2seq (our method):A *seq2seq* model with the 2-layer attention strategy [25] that uses the question-worthy sentences identified by our proposed sentence selection method from the paragraph of the input sentence as the question-worthy context.

We chose LexRank and ConceptMax as an alternative context selection method to compare with our method because they can identify important sentences better than other strategies evaluated in [25]. It can be seen from Tables 3 and 4, FS-SM-seq2seq outperform other compared methods on all metrics on the SQuAD data set and on most metrics on the Car Manuals data set.

Table 3. Question generation evaluation on car manuals on SQuAD

Model (SQuAD)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Vanilla seq2seq	31.34	13.79	7.36	4.26	29.75
Transformer	37.528	18.097	9.457	5.0143	26.600
ConceptMax-seq2seq	41.700	16.551	8.205	4.099	28.772
LexRank-seq2seq	41.057	17.168	8.494	4.099	28.055
Para-seq2seq	33.152	13.786	06.585	03.2867	27.6876
FS-SM-seq2seq	**43.27**	**18.86**	**9.00**	**4.48**	**30.58**

Table 4. Question generation evaluation on car manuals

Model (Car manual)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Vanilla seq2seq	34.6012	16.5057	**10.11052**	**6.598**	28.1247
Transformer	28.1243	11.5928	6.3074	3.4219	25.2176
ConceptMax-seq2seq	35.2702	14.7947	9.3679	5.2965	26.9364
LexRank-seq2seq	35.4368	0.1601	9.764	6.1662	28.08
Para-seq2seq	35.13123	15.97419	9.2094	5.5600	28.0959
FS-SM-seq2seq	**36.9870**	**17.6561**	9.7696	5.41238	**29.5423**

5 Conclusion and Future Work

We presented a method for selecting question-worthy sentences from a text passage and using these sentences as contexts for question generation. For identifying question-worthy sentences, a feature-based method is designed based on context-based and sentence-based features. A 2-layer attention strategy is applied to incorporate the question-worthy context into a *seq2seq* model. Experimental results showed that using the question-worthy context for question generation *seq2seq* models have achieved better results than baselines on both Car Manuals and SQuAD datasets.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.0473
2. Barandiaran, I.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell **20**(8), 1–22 (1998)
3. Brin, Sergey, Page, Lawrence: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30**, 107– (01 1998)
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. Danon, G., Last, M.: A syntactic approach to domain-specific automatic question generation. arXiv preprint arXiv:1712.09827 (2017)

6. Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1342–1352 (2017)
7. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research **22**, 457–479 (2004)
8. Fabish, A.: MS Windows NT kernel description, https://github.com/adamfabish/Reduction
9. Galanis, D., Lampouras, G., Androutsopoulos, I.: Extractive multi-document summarization with integer linear programming and support vector regression. In: Proceedings of COLING 2012. pp. 911–926 (2012)
10. Gülçehre, Ç., Ahn, S., Nallapati, R., Zhou, B., Bengio, Y.: Pointing the unknown words. CoRR **abs/1603.08148** (2016), http://arxiv.org/abs/1603.08148
11. Gupta, S., Nenkova, A., Jurafsky, D.: Measuring importance and query relevance in topic-focused multi-document summarization. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 193–196. Association for Computational Linguistics (2007)
12. Heilman, M., Smith, N.A.: Good question! statistical ranking for question generation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 609–617. Association for Computational Linguistics, Los Angeles, California (Jun 2010), https://www.aclweb.org/anthology/N10-1086
13. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
14. Klein, G., Kim, Y., Deng, Y., Crego, J.M., Senellart, J., Rush, A.M.: Opennmt: Open-source toolkit for neural machine translation. CoRR **abs/1709.03815** (2017), http://arxiv.org/abs/1709.03815
15. Li, S., Ouyang, Y., Wang, W., Sun, B.: Multi-document summarization using support vector regression. In: Proceedings of DUC. Citeseer (2007)
16. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain (Jul 2004)
17. Lindberg, D., Popowich, F., Nesbit, J., Winne, P.: Generating natural language questions to support learning on-line. In: Proceedings of the 14th European Workshop on Natural Language Generation. pp. 105–114 (2013)
18. Mazidi, K., Nielsen, R.D.: Linguistic considerations in automatic question generation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 321–326 (2014)
19. Mazidi, K., Nielsen, R.D.: Leveraging multiple views of text for automatic question generation. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) Artificial Intelligence in Education. pp. 257–266. Springer International Publishing, Cham (2015)
20. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp. 404–411 (2004)
21. Mitkov, R., Ha, L.A.: Computer-aided generation of multiple-choice tests. In: Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing (2003)
22. 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 of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002)

23. Patil, N.R., Patnaik, G.K.: Automatic text summarization with statistical, linguistic and cohesion features. In: International Journal of Computer Science and Information Technologies (2017)
24. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100, 000+ questions for machine comprehension of text. CoRR **abs/1606.05250** (2016), http://arxiv.org/abs/1606.05250
25. Ren, G., Ni, X., Malik, M., Ke, Q.: Conversational query understanding using sequence to sequence modeling. In: Proceedings of the 2018 World Wide Web Conference. pp. 1715–1724. International World Wide Web Conferences Steering Committee (2018)
26. Ren, P., Wei, F., Zhumin, C., Jun, M., Zhou, M.: A redundancy-aware sentence regression framework for extractive summarization. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 33–43 (2016)
27. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. CoRR **abs/1704.04368** (2017), http://arxiv.org/abs/1704.04368
28. Sharma, S., El Asri, L., Schulz, H., Zumer, J.: Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. CoRR **abs/1706.09799** (2017), http://arxiv.org/abs/1706.09799
29. Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., Wang, S.: Answer-focused and position-aware neural question generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://www.aclweb.org/anthology/D18-1427
30. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. Advances in NIPS (2014)
31. Vanderwende, L., Suzuki, H., Brockett, C., Nenkova, A.: Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management **43**(6), 1606–1618 (2007)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
33. Yao, K., Zhang, L., Luo, T., Tao, L., Wu, Y.: Teaching machines to ask questions. In: IJCAI. pp. 4546–4552 (2018)
34. Yuan, X., Wang, T., Gulcehre, C., Sordoni, A., Bachman, P., Zhang, S., Subramanian, S., Trischler, A.: Machine comprehension by text-to-text neural question generation. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. pp. 15–25. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/W17-2603, https://www.aclweb.org/anthology/W17-2603
35. Yuan, X., Wang, T., Trischler, A.P., Subramanian, S.: Neural models for key phrase detection and question generation (Feb 7 2019), uS Patent App. 15/667,911
36. Zhao, Y., Ni, X., Ding, Y., Ke, Q.: Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3901–3910 (2018)
37. Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., Zhou, M.: Neural question generation from text: A preliminary study. CoRR **abs/1704.01792** (2017), http://arxiv.org/abs/1704.01792