Doshisha University, Universität zu Lübeck and German Research Center for Artificial Intelligence at TRECVID 2023: MIQG Task

Zihao Chen[1], Frédéric Li[2], Marc S. Seibel[2], Nele S. Brügge[3], Miho Ohsaki[1], Heinz Handels[2,3], Marcin Grzegorzek[2], Kimiaki Shirahama[1]
[1]Doshisha University, 1-3, Tatara Miyakodani, Kyotanabe, 610-0394 Kyoto, Japan
[2]Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
[3]German Research Center for Artificial Intelligence (DFKI), 23562 Lübeck, Germany
E-mail: zihaochenyichang@gmail.com

***Abstract* – This paper presents the approaches proposed by the *doshisha_uzl* team to address the Medical Instructional Question Generation (MIQG) task of TRECVID 2023. Given a clip from a video containing medical content and its associated transcript, we explored various text-to-text approaches to generate relevant medical questions using the *Flan-T5* language model [3]. Our approaches are based on firstly generating a text summary of the clip contents by leveraging both video and associated text transcript modalities using the *mPLUG* multi-modal text generation model [9]. Secondly, the generated summaries are sent as input of Flan-T5 that is fine-tuned to output relevant questions on the MIQG train set. We also experimented with variants of this baseline approach using various data augmentation techniques on the mPLUG summaries and ground truth questions of the train set, and/or extracting keywords from the mPLUG summaries to feed Flan-T5 with information akin to an answer-span. Our experiments show that Flan-T5 performs the best for the MIQG task when trained using the mPLUG summaries without augmentation and with keywords obtained by the *Topical Page Rank* method [7] as input. This approach yields a BLEU / BLEU-4 / ROUGE-2 / ROUGE-L / BERTScore of 0.15828 / 0.05153 / 0.27752 / 0.47825 / 0.91092 respectively, achieving overall top BLEU, BLEU-4 and ROUGE-2 scores in the MIQG challenge of this year, and second best for ROUGE-L and BERTScore.**

***Keywords* – natural language processing, answer-aware question generation, keyword extraction, video summarisation**

I. Introduction

The proliferation of online videos has revolutionised the way we access and disseminate information. Instructional videos have become a preferred medium for imparting knowledge and skills, offering an effective and efficient step-by-step guide. In the field of healthcare, medical educational videos hold immense potential. They can provide important information through a combination of visual demonstrations and verbal explanations, answering questions of healthcare consumers.

In light of this ongoing change, the TRECVID 2023 challenge [1] aims to set the challenge of answering medical video questions. Its main objective is to promote research to develop systems capable of understanding medical videos and providing visual answers to natural language questions. It also aims to equip these systems with multimodal capabilities that enable them to generate instructive questions based on the content of medical videos and their transcripts.

Building on the success of the first MedVidQA collaborative task at the BioNLP workshop during ACL 2022, MedVidQA 2023 is expanding its horizons at TRECVID. The challenge includes two tasks: Video Corpus Visual Answer Localisation (VCVAL) and Medical Instructional Question Generation (MIQG).

Our approach addresses MIQG (task 2) combining state-of-the-art approaches. MIQG is a text generation task that aims to generate an instructional question given a video clip containing medical-related instructions and its associated transcript.

We use the cross-modal vision-language foundation model mPLUG Owl [10] to generate summaries and incorporate the information from both video and transcripts. The model provides a comprehensive world knowledge, as it has been pre-trained on a large amount of data and achieves state-of-the-art results in many vision language downstream tasks like video and image captioning or visual question answering. For question generation, we finetune the Text-To-Text Transfer Transformer FLAN-T5 on the mPLUG predictions and the corresponding ground truth questions. We also investigated the influence of decoding algorithms, augmentation techniques and keyword extraction on the performance of the model. We found that the best performing approach was achieved with beam search decoding for the Flan-T5 output selection, without text augmentation and with additional keyword extraction.

II. Methodology

A. Proposed approach

Our proposed approach for the MIQG task aims to extract relevant information for the medical question generation process jointly using both video and text modalities via the video clips and their associated transcripts. This objective is motivated by our observation of the provided dataset samples that suggested that either video or transcripts used separately may be insufficient on their own. The most relevant content of some examples was for instance contained in the video only (e.g. missing or flawed text transcript, non-English transcript but English instructions written in the video, etc.), while for others the text transcript contained information more

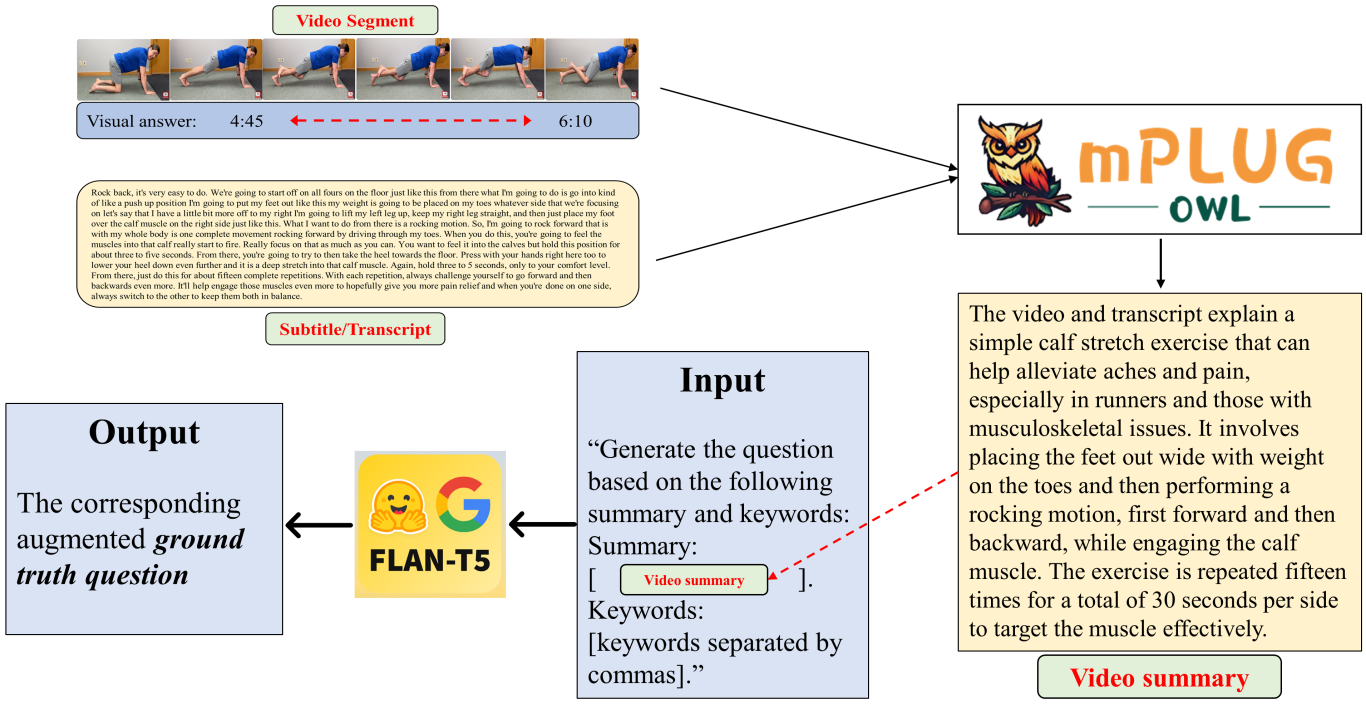

Fig. 1. An overview of our proposed approach. The video extracts and associated transcripts are first sent to a pre-trained mPLUG-Owl model with a prompt asking it to generate a summary of the video clip contents. The summaries are then used to fine-tune a Flan-T5 model for question generation.

susceptible to be useful for the generation task.

Fig. 1 shows an overview of our proposed approach. More specifically, the latter is based on leveraging both multimodal (video and text) and unimodal (text) language models existing from the literature. We first used the cross-modal vision-language foundation model mPLUG-Owl [10] that generates summaries of the video clip contents using the video, its associated text transcript and a prompt as input. The generated summaries were then sent to the text-to-text Flan-T5 transformer model [3] (Flan-T5-base variant) with the input prompt: *"Generate the question based on the following summary: [mPLUG summary]"*. We refer to the aforementioned approach as *Baseline*. Each of its components can be described as follows:

1) Extraction of mPLUG summaries

Pre-trained mPLUG weights for text and video processing were obtained from *HuggingFace* (mplug-owl-llama-7b-video)[1]. We passed each video clip and the corresponding transcript to the mPLUG model using the prompt described in Fig. 2. We used the default parameters, i.e. for text generation we used top-k sampling with k=5 and a maximum of 512 tokens. From each video clip, four equidistantly spaced images were processed by the mPLUG model.

```
 The following is a conversation
between a curious human and AI
assistant.
Human:  Here is a video and a
transcript of someone explaining the
video.
Human:  <|video|>
Human:  Transcript:  {transcript}
Human:  Summarize the instructions
provided by the video and the
transcript in less than {num_sentences}
sentences.
AI:
```

Fig. 2. Input prompt provided to the mPLUG model.

[1]https://huggingface.co/MAGAer13/mplug-owl-llama-7b-video

2) Question generation with FLAN-T5

We used a pre-trained Flan-T5 transformer model [3] as backbone model to generate medical instructional questions. The Flan-T5 model achieves state-of-the-art performances on several evaluation benchmarks and can be directly used for unseen zero-shot, few-shot prompting, as it has been finetuned on both with and without exemplars (i.e., zero-shot and few-shot) and with and without chain-of-thought with 1.8K multi-tasks phrased as instructions. The model was obtained from *HuggingFace* (flan-t5-base)[2]. The publicly released Flan-T5 checkpoints are initialised from the T5 1.1 LM-Adapted checkpoints and instruction-finetuned.

To get better generation quality for our specific tasks, we decided to fine-tune the pre-trained Flan-T5 model. In the fine-tuning process, the generated summaries obtained by the mPLUG-Owl model were sent to the pre-trained Flan-T5 model in addition to the following input instruction prompt: *"Generate the question based on the following summary: [mPLUG summary]"*. The corresponding ground truth questions were used as targets for the output to reproduce the similar fine-tuning procedure data format described in the official paper. We used this fine-tuned Flan-T5 model and applied the *Generation* function of *HuggingFace* to generate the medical instructional questions. More specifically, we chose the default parameters with two prediction strategies, the beam search decoding [4] and the nucleus sampling decoding [5], to generate texts with the top 5 sentences and a maximum of 32 tokens following the same input format as used in the fine-tuning procedure. Compared with the beam search method, the nucleus sampling method can decode more fancy and diverse sentences, but the beam search method can generate clearer sentences.

B. Variations of the Baseline approach

To attempt to improve the Baseline approach, we experimented with alternative strategies to fine-tune the Flan-T5 model for the MIQG task. We in particular investigated two strategies, with the first one being based on training data augmentation (on both mPLUG summaries and ground truth questions), and the second one based on keyword extraction from the mPLUG summaries to provide additional information to the model during the fine-tuning of Flan-T5.

1) Data augmentation

Data augmentation was applied on both the mPLUG summaries and the ground truth questions to increase the size of the training set. Different strategies were employed depending on whether the augmentation was performed on the summaries or questions.

To augment the summaries, we asked mPLUG-Owl to generate different summaries of the same sample by changing the input prompt. Requests to create summaries of different lengths n were made using the prompt: "Summarize the instructions provided by the video and the transcript in less than n sentences" with $n \in \{1,3\}$. Taking advantage of the fact that mPLUG-Owl is non-deterministic, four augmented samples by getting two summaries with $n = 1$ and two others with $n = 3$. It should however be noted that mPLUG-Owl occasionally ignored the length restriction, sometimes yielding summaries longer than the maximum length specified in the prompt.

To augment the ground truth questions, we experimented with methods from the *TextAugment* Python library [8]. More specifically, the Word2Vec (random words replaced with similar ones determined using the Word2vec embeddings), WordNet (same as Word2Vec, except that only nouns and/or verbs can be replaced, and WordNet embeddings are used), Translate (text translated into French, then back to English), double Translate (consecutive translation loops with English/French and English/German) and synonym replacement with Easy Data Augmentation (EDA - random words replaced with a random synonym) methods were tested. A manual subjective check of the augmented questions on some samples of the validation set was performed to find the method(s) leading to the most natural augmentation. It was observed that Word2Vec often led to unnatural questions and vocabulary. Both translation methods were often deterministic, leading to redundant augmented questions. WordNet performed better despite sometimes generating questions identical to the original ground truth. EDA also yielded results that were deemed satisfactory, despite sometimes leading to sentences whose meaning departed from the original question. Based on the aforementioned observations, it was decided to augment each question two times with Word2Vec, and two times with EDA, leading to up to four augmented samples per original question. Potential duplicates were removed to avoid redundancy in the training set.

The augmentation on the mPLUG summaries was combined with the augmentation on the ground truth questions. It should however be noted that we found out a mistake in the runs that we submitted using augmentation: the original mPLUG summaries generated with $n = 5$ were mistakenly omitted, and only the ones with $n = 1$ and $n = 3$ were used to fine-tune Flan-T5. This augmentation led to a total of $43,013$ training samples, up from $2,710$ originally.

2) Keyword extraction

The idea to extract keywords was attempted after checking the question generation literature that differentiates two main families of approaches [11]. The first consists in *answer-aware* approaches where the model trained to generate the question based on some input text is also fed with "elements of answer" to guide the generation process. Such elements may take various forms such as word(s) from the input text related to the target question, also referred to as *answer-span*, or even answers directly. The second category is referred to as *answer-agnostic* methods where the question generation is performed by the model in a completely unguided way. Because answer-agnostic question generation is a one-to-many problem (i.e. several outputs may be suitable for a given input text), it is hypothesised to be less effective than answer-aware methods in practice [11].

In the light of this observation, we hypothesise that keywords

[2]https://huggingface.co/google/flan-t5-base

from the mPLUG summaries may help the generation model to output questions more relevant to the core content of the input text, thus acting as helpful answer-spans during the training of the model. We therefore investigated keyword extraction methods. For this purpose, we used the *Python Keyphrase Extraction* (PKE) library [2] that implements various statistical-, graph- or feature-based keyword extraction methods proposed in the literature. We arbitrarily chose the number of keywords to extract to be three, and applied them on the mPLUG summaries of the validation set. A subjective evaluation of the methods was carried out to determine the best performing one in terms of keyword relevance to the question associated to the summary. After this analysis, it was chosen to use the *Topical Page Rank* approach [7] with a number of extracted keywords set to three. When keywords were used, the Flan-T5 input prompt was modified to "Generate the question based on the following summary and keywords: Summary: [mPLUG summary]. Keywords: [keywords separated by commas].".

3) Submitted runs

All eight configurations involving the Baseline with beam or nucleus search, with or without data augmentation and with or without keyword extraction were tested. More specifically, each configuration was used to train one Flan-T5 model subsequently used to generate questions on the provided MIQG testing set. The 80 generated questions were then subjectively evaluated to check how sound they looked like. The five best configurations were then determined and submitted as runs 1 to 5 ordered from most to least promising. In details, each run uses the Baseline described in Section II.A. with the parameters described in Table 1.

Run	Search strategy	Augmentation	Keywords
run-1	beam	no	yes
run-2	beam	yes	no
run-3	beam	no	no
run-4	beam	yes	yes
run-5	nucleus	no	no

Table 1: Configuration used with the Baseline for each submitted run.

III. Results

The results of each submitted run on the testing set of the MIQG task are provided in Table 2.
It can be seen that the obtained performances mostly correlate with our ranking of the runs, with the exception that run-3 is arguably better than run-2 since the former outperforms the latter for all metrics except BLEU-4. Our run-1 achieves top 1 BLEU, BLEU-4, ROUGE-2 scores, and top 2 ROUGE-L and BertScore for the MIQG task of TRECVID 2023. Additionally, our run-3 achieves top 1 BertScore overall.

IV. Discussion

Dataset A first look at the examples contained in the provided dataset hinted at the fact that neither using only the video nor the text transcript could lead to optimal performances. It was for instance found out that in some cases, videos lacked descriptive transcripts (e.g. transcript unrelated to the actual medical content, not in English, etc.), or contained only instructions through on-screen text or actions. On the other hand, the dataset encompasses a diverse range of video types, including animation-style videos which may impact the quality of video feature extraction because these videos exhibit distinct visual and auditory characteristics compared to real-life videos. These observations led us to seek an approach that could extract relevant information from both video and text modalities. Finally, it can be mentioned that a small portion of the dataset was unavailable due to YouTube videos being set to private or removed entirely. This could have an influence on the robustness of the trained models by lowering the diversity of the training set.

Methods Two main aspects regarding our proposed methods can be highlighted: the impact of data augmentation and of the model output selection strategy.
While data augmentation can typically be beneficial to train a model, we observed it could lead to a degradation of performances compared to not using any in our case. This is hypothesised to be due to two factors: the first is that it was observed that the simple augmentation techniques of the *TextAugment* library could sometimes lead to augmented questions whose meaning departed from the original ground truth. Because of the lack of time, augmentations performed separately either on the summaries or the questions were not tested, but should be in future investigative studies. The second is that forcing mPLUG to generate several variations of the same summary could lead to the extraction of sub-optimal keywords as exemplified in Figure 3. This could explain the relatively mediocre performances of *run-4* that combines augmentation and keyword extraction.
It was also observed that the output selection method for the Flan-T5 model was a critical factor in the quality of the generated questions. Our study found that beam search consistently outperforms nucleus search, as exemplified by the top four configurations (out of eight) using beam search, while the bottom four use nucleus search. In our studies, only beam and nucleus search were tested as the two most popular output selection methods, but more could be investigated in the future.

Limitations There are some limitations and areas for potential improvement that should be acknowledged. Firstly as previously mentioned, the mistake performed with the augmentation should be corrected by adding back the mPLUG summaries generated with $n = 5$ or less sentences to the training set. Additionally, the individual impact of the data augmentation should be tested separately for both summaries and questions. Secondly, hyperparameters such as the number of keywords, "k" for top-k sampling and adapting to video length and the size of mPLUG summaries in the models were

Approach	BLEU	BLEU-4	ROUGE-2	ROUGE-L	BERTScore
run-1	**0.15828**	**0.05153**	**0.27752**	0.47825	0.91092
run-2	0.14352	0.04546	0.24478	0.44685	0.90523
run-3	0.14593	0.03875	0.27339	0.47439	**0.91099**
run-4	0.13289	0.03404	0.24421	0.45659	0.90780
run-5	0.09300	0.01627	0.20023	0.40716	0.90248
Overall min	0	0	0.12262	0.26083	0.85332
Overall mean	0.10041	0.02867	0.22098	0.41484	0.89717
Overall max	**0.15828**	**0.05153**	**0.27752**	**0.47924**	**0.91099**

Table 2: Performance metrics of the runs submitted by the *doshisha_uzl* team to the MIQG task. Global metrics are provided at the bottom.

mPLUG prompt: "Summarize the instructions provided by the video and the transcript in less than n sentences."

Output for sample 1 of the training set:

$n = 5$: "In the video, a physical therapist explains the Epley maneuver [...]. Both involve a series of head movements [...]."
Keywords: "head movements", "epley maneuver", "movement"

$n = 3$: "The instructions for performing the Epley maneuver to treat vertigo involve a patient turning their head [...]. After 30 seconds, they roll over onto their left side, tilt their head down towards their left shoulder, and maintain this position for 30 seconds, [...]."
Keywords: "left side", "head", "seconds"

$n = 1$: "The video and transcript describe two Epley exercises for treating benign paroxysmal positional vertigo (BPPV). The first exercise involves [...], and the second exercise requires them to [...]."
Keywords: "second exercise", "first exercise", "epley exercises"

Fig. 3. Augmented mPLUG summaries for sample 1 of the training set, and their extracted keywords with the Topical Page Rank method. Despite the different summaries being close to each other from the content point of view, the most relevant keywords extracted with Topical Page Rank were obtained for $n = 5$.

arbitrarily chosen due to time constraints, and would benefit to be selected in a more systematic manner. Finally, the quality of keyword extraction could be enhanced to ensure that extracted keywords closely relate to the video content, thus improving their effectiveness in guiding question generation. This could for instance be done by experimenting with fine-tuning large language models for this task or associated ones (e.g. answer-span detection) using a benchmark dataset where keywords would be available for the training of the model, such as the Natural Questions dataset [6].

V. Conclusion

The approaches submitted by the doshisha_uzl team to the MIQG challenge of TRECVID 2023 were presented in this paper. They leverage large language models from the literature by first asking the mPLUG text generation model to summarise the contents of the video clip using both video and text as input modalities, and then using the Flan-T5 model to generate a question using the mPLUG summary as input. Variations of this baseline involving either text augmentation and/or keyword extraction were tested as well. Our results show that keyword extraction without data augmentation improves the quality of the generated questions. Our *run-1* submission achieves either top 1 or top 2 in all metrics of the challenge of this year.

Due to time constraints, many aspects of the proposed approaches could not be investigated in depth. Improvements regarding testing augmentation on the questions and summaries separately, and devising a more systematic strategy for hyper-parameter selection should be investigated. Other future work will also focus on exploring learning-based keyword extraction methods instead of the graph-based approaches that were tested and used in the frame of this study.

REFERENCES

[1] George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Eliot Godard, Lukas Diduch, Yvette Graham, , and Georges Quénot. Trecvid 2023 - a series of evaluation tracks in video understanding. In *Proceedings of TRECVID 2023*. NIST, USA, 2023.

[2] Florian Boudin. pke: an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016*, pages 69–73, Osaka, Japan, December 2016.

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. arXiv pre-print, 2022. arXiv:2210.11416.

[4] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, 2017.

[5] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[6] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguists*, 7:453–466, 2019.

[7] Z. Liu, W. Huang, Y. Zheng, and M. Sun. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[8] Vukosi Marivate and Tshephisho Sefara. Improving short text classification through global augmentation methods. In *Proceedings of CD-MAKE 2020*, 2020.

[9] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mPLUG-2: A modularized multi-modal foundation model across text, image and video. In *Proceedings of ICML*, 2023.

[10] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

[11] R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng. A review on question generation from natural language text. *ACM Transactions on Information Systems*, 40(14):1–43, 2021.