Non-Essential is NEcessary: Order-agnostic Multi-hop Question Generation

[1]Kyungho Kim, [1]Seongmin Park, [2]Junseo Lee, [1]Jihwa Lee
[1]ActionPower, [2]Boston University
[1]{kyungho.kim, Seongmin.park, Jihwa.lee}@actionpower.kr, [2]jslee011@bu.edu

Abstract

Existing multi-hop question generation (QG) methods treat answer-irrelevant documents as *non-essential* and remove them as impurities. However, this approach can create a training-inference discrepancy when impurities cannot be completely removed, which can lead to a decrease in model performance. To overcome this problem, we propose an auxiliary task, called **order-agnostic**, which leverages non-essential data in the training phase to create a robust model and extract the consistent embeddings in real-world inference environments. Additionally, we use a single language model (LM) to perform both **ranker** and **generator** through a prompt-based approach without applying additional external modules. Furthermore, we discover that appropriate utilization of the *non-essential* components can achieve a significant performance increase. Finally, experiments conducted on HotpotQA dataset achieve state-of-the-art.

Keywords: multi-hop question generation, order-agnostic, multi-task learning

1. Introduction

Question Generation (QG) (Zhou et al., 2017; Du et al., 2017; Dong et al., 2019) is the task of automatically generating questions from a given input context and answer. With the increasing demand for automated question generation in various applications, such as education (Liu et al., 2012; Tsai et al., 2021), information retrieval (Zhao et al., 2011), and conversational agents (Gu et al., 2021), QG has become a prominent research area in Natural Language Processing (NLP).

Currently popular QG models mostly focus on generating questions in a single-hop manner, using a SQuAD-style QA dataset (Rajpurkar et al., 2016). Single-hop QG (Xiao et al., 2020; Ushio et al., 2022a) generates questions from a designated answer word by leveraging only a single sentence or phrase from the input context. Although useful in certain scenarios, accuracy of single-hop QG degrades when relevant contexts, which have intricate dependencies, are spread out across multiple regions in the input document.

Effectively applying QG in the real-world requires multi-hop reasoning, in which the model attends to necessary contexts spread out among non-adjacent documents. QG using dialogue transcripts, for instance, require understanding interactions across speaker boundaries (Jeong et al., 2020). In video modality, multiple segments are extracted from video (Kim et al., 2020).

To overcome the limitation of single-hop QG, multi-hop QG systems typically integrate a *ranker* that ranks and extracts only answer-relevant context segments. In addition, many previous research on multi-hop QG (Su et al., 2020a; Sachan et al., 2020; Pan et al., 2020; Xie et al., 2020; Su et al., 2022; Zhang and Bansal, 2019) employ additional external module to support the ranker as their context filtering modules to build document structure and complex relationships between sentences.

However, using an external module as the ranker to extract relevant documents can cause additional brittleness to the QG pipeline in real-inference scenario. One example is the training-inference discrepancy that originates from teacher-forcing the ranker during training. In practice, imperfect context result of the ranker degrades inference performance of the final QG model by relaying unstable extracted context, which not only has golden answers but also unnecessary information.

This paper proposes ***NENE***, a multi-hop QG model designed to be robust against noise from imperfect context extraction. The key idea of NENE is to use a *single* sequence-to-sequence LM that acts as both the **context ranker** and **question generator** through prompt-based approach, and utilize **order-agnostic encoder** as *auxiliary task* to minimize the impact of training-inference discrepancy. The sub-task, context ranker, introduces an objective to sort documents by relevant to the target question. The auxiliary task, order-agnostic encoder, enables the LM encoder to produce consistent context embeddings even in realistic inference scenarios where the ranker fails to arrange documents in order of relevance. During the process, NENE learns to incorporate parts of documents previous QG systems dismissed as "unnecessary" for better question generation. We make a surprising observation where bootstrapping the encoder for robustness enables the system to leverage information in seemingly irrelevant contexts to increase question generation quality.

LREC-COLING 2024, pages 12300–12306
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Our method is validated on the widely-used HotpotQA benchmark with various metrics. While achieving state-of-the-art performance in multi-hop QG, we train only a single LM along with order-agnostic properties by multi-task learning.

Our main contributions are as follows:

- We propose a novel framework, **NENE**, that utilizes a single language model as both the **ranker** and the **generator** using a prompt without any additional modules.
- To help the model learn the dependency between documents in an implicit way and tackle the training-inference gap, we introduce an auxiliary task called **order-agnostic encoder**, which ensures that the LM encoder generates consistent embeddings regardless of the order and amount of answer-irrelevant documents.
- We evaluate our method on HotpotQA benchmarks and achieve state-of-the-art performance, using widely used metrics such as BLEU, ROUGE, and METEOR scores.

2. Preliminary

2.1. Multi-hop Question Generation

The objective of the Multi-hop QG task is to generate a question Q with a given answer a and given text corpus of M documents $\mathbb{D} = \{D_1, ..., D_M\}$. This can be represented as follows:

$$Q = f_{qg}(a, \mathbb{D}) \tag{1}$$

where f_{qg} refers to the question generation model and only a subset of documents $\mathbb{S} \subset \mathbb{D}$ is relevant to the question and answer.

Also, there are many recent studies on multi-hop QG utilizing the additional module. MulQG (Su et al., 2020a) and GATE (Sachan et al., 2020) use an entity graph to link sentences that mention the same object. SGGDQ (Pan et al., 2020) builds DP-based semantic graphs to capture global structures of documents and facilitate reasoning. FRA (Xie et al., 2020) applies reinforcement learning by designing discriminators for QG-specific rewards. QA4QG (Su et al., 2022) and semQG (Zhang and Bansal, 2019) add additional QA modules for multi-hop QG.

3. Methodology

In this section, we introduce our architecture to train multi-tasks with a single LM for robust multihop-QG. Figure 1 presents an overview of our approach, which introduces a ranker as a subtask (Section 3.1), order-agnostic encoder as an auxiliary task (Section 3.2), and a final multi-hop QG module as a main task (Section 3.3).

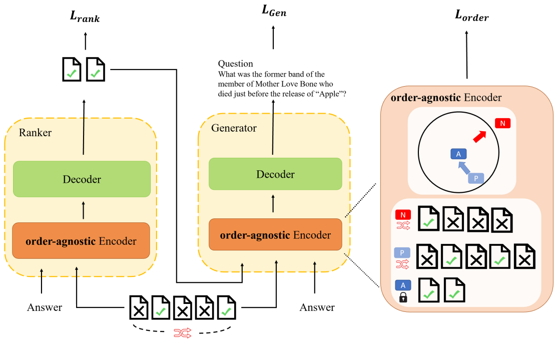

Figure 1: Overall structure of NENE. A, P, and N respectively denote the anchor, positive, and negative.

3.1. Subtask: Ranker

Among a set of documents entering the input of a multi-hop QG, answer-irrelevant documents, called *non-essential*, exist that are unnecessary to generate a question. The sub-task is designed to reflect the above data characteristics. The ranker module extracts answer-relevant documents, called *essential*, required to generate questions by referring to the answer from a given corpus. The result of the ranker module can be utilized to provide refined data to the generator and to improve the generation ability during inference time. Additionally, "select relevant document indexes" is used as the task definition in the prompt phrase for a given sub-task to improve clarity and performance. The final input form of the ranker is [select relevant document indexes: answer: a, context: $\mathbb{D}$]. The output of ranker is the index of answer-relevant document which is produced by generation manner. It can be denoted as follows:

$$S = f_{rank}(a, \mathbb{D}) \tag{2}$$

where S means the answer-relevant subset of documents and f_{rank} means the ranker model including the prompt phase. With the training, the loss of the ranker using cross-entropy is defined as follows:

$$L_{rank} = -\Sigma s \log(\hat{s}) \tag{3}$$

where $s \in \mathbb{S}$ means index of answer-relevant documents and $\hat{s}$ is the predicted probability of s.

3.2. Auxiliary Task: Order-agnostic Encoder

The auxiliary task is designed to help generate appropriate questions in any given situation without additional external modules, even if the ranker module fails to get the correct golden output for a given text corpus. In other words, regardless of the order and number of unrelated documents (*non-essential*), the encoder embeddings should always

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
with Golden Supporting Facts Sentences						
ASs2s-a (Kim et al., 2018)	37.67	23.79	17.21	12.59	17.45	33.21
SemQG (Zhang and Bansal, 2019)	39.92	26.73	18.73	14.71	19.29	35.63
F+R+A (Xie et al., 2020)	37.97	-	-	15.41	19.61	35.12
SGGDQ (DP) (Pan et al., 2020)	40.55	27.21	20.13	15.53	20.15	36.94
ADDQG (Wang et al., 2020)	44.34	31.32	22.68	17.54	20.56	38.09
QA4QG (LARGE) (Su et al., 2022)	49.55	37.91	30.79	25.70	27.44	46.48
NENE (BASE)	**49.62**	**40.1**	**32.48**	**26.22**	**37.22**	**49.14**
Full Document Context *(w/o Golden Supporting Facts Sentences)*						
MulQG (Su et al., 2020b)	40.15	26.71	19.73	15.2	20.51	35.3
GATE$_{NLL+CT}$ (*Sachanet al.*, 2020)	-	-	-	20.02	22.40	39.49
LowResourceQG (Yu et al., 2020)	-	-	-	19.07	19.16	39.41
QA4QG (BASE) (Su et al., 2022)	43.72	31.54	24.47	19.68	24.55	40.44
QA4QG (LARGE) (Su et al., 2022)	**46.45**	**33.83**	**26.35**	**21.21**	25.53	42.44
NENE (BASE)	44.40	33.46	25.14	19.01	**32.93**	**42.61**

Table 1: Performance comparison between NENE and previous MQG methods on the HotpotQA dataset. The results, except ours, are reported from (Su et al., 2022). **Bold** represents the best performance among all experiments. Underline indicates the best score excluding the LARGE model settings.

be the same when all answer-relevant documents are included in the input. We define the above intuition using the relationship between an anchor and a positive example. An anchor consists of only essential documents, and positive contains all essential documents, but non-essential documents are also randomly mixed. Additionally, negative samples are designed to prevent all embeddings from becoming similar to each other. An easy negative sample is the input corpus consisting of only non-essential documents. A hard negative sample includes parts of answer-relevant documents as input. This negative sample prevents it from becoming more similar to anchor embedding at a certain level. As a result, the model is able to learn the relationships between sentences and documents in an unsupervised manner without the assistance of external modules. To implement this, the triplet margin loss is adopted. It can be denoted as follows:

$$L_{order} = \max(0, d(A, P) - d(A, N) + \gamma) \quad (4)$$

where d means the distance between encoder embedding of inputs, A means anchor which consists of only all answer-relevant documents, P means positive sample, N means negative sample, and γ is the margin parameter controlling the minimum relative distance.

3.3. Main Task: Generator

The generator module receives the original data and output of ranker as input and performs multi-hop QG. Except for using a different task definition in the prompt phase, the generator model is identical to the one used in the ranker module. For QG, the task definition “generate question” is used as the prompt. The final input form of the generator module is [generate question: answer: a, context: S, $\mathbb{D}$]. It can be denoted as follows:

$$Q = f_{gen}(a, S, \mathbb{D}) \quad (5)$$

where f_{gen} means the final generator module for multi-hop QG. The loss of the generator is computed by cross-entropy, which is expressed as:

$$L_{gen} = -\Sigma q \log(\hat{q}) \quad (6)$$

where $q \in \mathbb{Q}$ refers to the each token of question and $\hat{q}$ is the predicted probability of that token.

Finally, we integrate all losses as follows:

$$L_{final} = L_{gen} + \alpha L_{rank} + \beta L_{order} \quad (7)$$

Input	BLEU	METEOR	ROUGE-L
Ranker	33.47	25.21	32.55
+ non-essential	43.29	31.90	41.80
+ order-agnostic	**44.40**	**19.01**	**42.61**

Table 2: Ablation study of generator.

Extractor	F1	EM
Anchor	20.59	21.47
+ Positive	72.40	41.18
+ Easy Negative	74.43	**44.25**
+ Hard Negative	**74.56**	**44.25**

Table 3: Ablation study of ranker.

4. Experiment

4.1. Experiment Setting

In order to evaluate the efficacy of NENE, we carry out experiments on the HotpotQA Yang et al. (2018) dataset. Also, we use the LMQG model (Ushio et al., 2022b) as a baseline model, pre-trained for single-hop question generation, based on T5-base (Raffel et al., 2019). We train the NENE framework with the AdamW optimizer (Loshchilov and Hutter, 2017) and cosine annealing with warm-up restart scheduler (Loshchilov and Hutter, 2016). The total epoch is set to 15, and the first 3 epochs are used as a warm-up session. The batch size is set to 16. The initial learning rate is set to 0.001, and the random seed is set to 10. We set γ to 1 for margin value and the cosine distance is used for triplet margin loss. We use the default cross-entropy loss for both the ranker loss and generator loss. The total loss is balanced with α and β where 1.0 and 0.1, respectively. In addition, to learn a order-agnostic property, we randomly shuffle input data in each iteration. The Golden Supporting Facts Sentences from HotpotQA are considered as all the necessary information for problem-solving. Additionally, non-essential sentences are utilized for negative samples.

4.2. Performance Evaluation

Table 1 shows the multi-hop QG performance of our proposed framework, NENE, along with that of baselines. The upper half of Table 1 evaluates our framework, assuming that the answer-relevant documents are known as oracle in advance. The lower half evaluates our framework when given the entire corpus without knowing which documents are related to the answer. Our proposed model outperforms all compared models in every evaluation metric when QG is performed using fused data augmentation with authentic answer-relevant documents, known as an oracle, without the ranker module during inference time. Additionally, in an end-to-end multi-hop QG scenario where a ranker module is used to extract answer-relevant documents followed by QG, our model achieves the best performance in all metrics except for BLEU-4 when having the same model size condition. Moreover, our proposed model shows better results in METEOR and ROUGE-L than the LARGE size model in full document context setting. This result indicates that the proposed subtask and auxiliary task are effective in performing the multi-hop QG main task.

4.3. Ablation Study of Generator

As shown in Table 2, we conduct an ablation study to validate the usefulness of each proposal. Each row include extra modules, including those mentioned in the preceding rows. We observe that all the proposed factors contribute to the improvement of QG performance. Note that excluding non-essential documents dramatically hinders performance. We find that including non-essential documents for multi-hop QG are important in learning the relationship between documents in an unsupervised and implicit manner without additional modules. In addition, the combination of non-essential and order-agnostic shows best performance, which highlights that proper utilization of non-essential documents is necessary for the multi-hop question generation model. There are two reasons for such a phenomenon. First, the ranker module may fail to accurately extract the essential documents during inference. In such cases, the QG module, which only uses the result of the ranker, has limited access to the information required for generating appropriate multi-hop questions. Second, the more complex input data is prepared by considering the order-agnostic property, which aims to ensure consistent results regardless of the composition of the data.

4.4. Ablation Study of Ranker

Table 3 shows the performance gain of ranker brought about by each component of the order-agnostic encoder. The anchor refers to the utilization of a conventional encoder without considering the order-agnostic property. The result demonstrates that the ranker module with an auxiliary task incorporating non-essential items leads to a significant performance improvement rather than using only answer-relevant documents. In other words, our unsupervised auxiliary task, designed to satisfy the order-agnostic property for the encoder that ensures similar embedding, extracted for inputs containing all answer-related elements regardless of their composition, helps to identify the relation of components and facilitates the ranker task. Additionally, our proposed hard negative achieves the best performance, indicating that the proper design for using non-essential can further improve performance.

5. Conclusion

In this paper, we propose a robust multi-hop QG framework, NENE, which fully utilizes the data previously thought to be non-essential based on a single LM through a prompt-based multi-task learning. To utilize the non-essential documents, we design subtask, auxiliary task, and main task of multi-hop QG by proposing the order-agnostic property, which guarantees consistent output and robust model by considering training-inference discrepancy. Through these modules, our model is able to extract the connotation from any situation, even if mandatory and non-essential information is misclassified, without the help of additional external model that explicitly define relationships between sentences. The proposed NENE is evaluated to prove effectiveness on the widely used multi-hop QG dataset, HotpotQA. The outcomes reveal a consistent and significant performance improvement over state-of-the-art models.

6. Limitations

While our model achieves a new state-of-the-art performance in multiple evaluation metrics, it still has several limitations. First, our framework trains an order-agnostic encoder with a set of all given documents as input. This strategy has the potential to exceed the max sequence length limit of model when the number of given documents is too large. Second, our order-agnostic task operates in an unsupervised manner with a simple structure, which suggests that our proposal is applicable to multiple tasks. However, this also means that it does not consider the characteristics of the main task. If appropriate guidance is given that reflects the properties of the main task, it is expected that the end-to-end performance of the model will be further improved.

7. Bibliographical References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, M. Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *ArXiv*, abs/1905.03197.

X. Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. Chaincqg: Flow-aware conversational question generation. *ArXiv*, abs/2102.02864.

Myeongho Jeong, Seungtaek Choi, Hojae Han, Kyungho Kim, and Seung won Hwang. 2020. Conditional response augmentation for dialogue using knowledge distillation. In *Interspeech*.

Kyungho Kim, Kyungjae Lee, and Seung won Hwang. 2020. Instructional video summarization using attentive knowledge grounding. In *ECML/PKDD*.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2018. Improving neural question generation using answer separation. In *AAAI Conference on Artificial Intelligence*.

Ming Liu, Rafael Alejandro Calvo, and Vasile Rus. 2012. G-asks: An intelligent automatic question generation system for academic writing support. *Dialogue Discourse*, 3:101–124.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

N. Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *ArXiv*, abs/1603.06059.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Annual Meeting of the Association for Computational Linguistics*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*.

Devendra Singh Sachan, Lingfei Wu, Mrinmaya Sachan, and William Hamilton. 2020. Stronger transformers for neural multi-hop question generation. *ArXiv*, abs/2010.11374.

Dan Su, Peng Xu, and Pascale Fung. 2022. Qa4qg: Using question answering to constrain multi-hop question generation. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8232–8236.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020a. Multi-hop question generation with graph convolutional network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4636–4647.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020b. Multi-hop question generation with graph convolutional network. *ArXiv*, abs/2010.09240.

Danny C.L. Tsai, Anna Y. Q. Huang, Owen H. T. Lu, and Stephen J. H. Yang. 2021. Automatic question generation for repeated testing to improve student learning outcome. *2021 International Conference on Advanced Learning Technologies (ICALT)*, pages 339–341.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022a. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asahi Ushio, Fernando Alva-Manchego, and José Camacho-Collados. 2022b. Generative language models for paragraph-level question generation. In *Conference on Empirical Methods in Natural Language Processing*.

Liuyin Wang, Zi-Han Xu, Zibo Lin, Haitao Zheng, and Ying Shen. 2020. Answer-driven deep question generation based on reinforcement learning. In *International Conference on Computational Linguistics*.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *International Joint Conference on Artificial Intelligence*.

Yuxi Xie, Liangming Pan, Dongzhe Wang, Min-Yen Kan, and Yansong Feng. 2020. Exploring question-specific rewards for generating deep questions. In *International Conference on Computational Linguistics*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jianxing Yu, Wei Liu, Shuang Qiu, Qinliang Su, Kai Wang, Xiaojun Quan, and Jian Yin. 2020. Low-resource generation of multi-hop reasoning questions. In *Annual Meeting of the Association for Computational Linguistics*.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Shiqi Zhao, Haifeng Wang, Chao Li, Ting Liu, and Yi Guan. 2011. Automatically generating questions from queries for community-based question answering. In *International Joint Conference on Natural Language Processing*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and M. Zhou. 2017. Neural question generation from text: A preliminary study.

In *Natural Language Processing and Chinese Computing*.

8. Language Resource References

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.