Cross-Domain Question Generation: A Comparative Study

Niloufar Beyranvand[1], **Aijun An**[1], **Heidar Davoudi**[2]
[1]Department of Electrical Engineering and Computer Science, York University, Canada
[2]Faculty of Science, Ontario Tech University, Canada
{nbeyran,aan}@yorku.ca
heidar.davoudi@ontariotechu.ca

1 Introduction

Question Generation (QG) from text has become increasingly popular, driven by its applicability in diverse areas such as educational reading comprehension tests (Chen et al., 2018; Kumar et al., 2018), enhancing datasets for question-answering system training (Sultan et al., 2020), and generating responses in conversational systems (Gu et al., 2021).

QG methods can be categorized into rule-based approaches and neural sequence-to-sequence (Seq2Seq) learning approaches. Rule-based methods use hand-crafted template rules based on linguistic features (such as semantic role labels) to convert sentences into questions (Chali and Hasan, 2015; Khullar et al., 2018). Although such rules are generally domain independent, they cannot capture the complexity or variety of ways humans ask questions and are also labor intensive to create.

In contrast, neural Seq2Seq models, particularly those based on RNNs and Transformers, have gained popularity due to their ability to model complex functions and extract effective features. Early Seq2Seq models used RNNs, such as LSTM with attention mechanisms (Zhao et al., 2018), while more recent advancements involve Transformer-based models like T5 (Raffel et al., 2020), BART (Lewis et al., 2020), and ProphetNet (Qi et al., 2020). However, Seq2Seq models require a large amount of labeled data to train and they generally perform well only in the domain where the training data originate. When there is a domain shift, their performance can decline significantly.

Recently, (Naeiji et al., 2023) proposed a QG method (which we name SRL-Seq2Seq) that takes advantage of both rule-based and Seq2Seq-based methods. It first converts the sentences and questions in the training data into their generalized semantic respresentations using semantic role labels (SRL), and then trains a Seq2Seq model based on these generalized representations to convert a SRL-label sentence into a SRL-labeled question. In the inference phase, an input sentence is first converted into its semantic representation with SRL, which is then converted into questions using the trained Seq2Seq model and an auxiliary function. The inference process is similar to rule-based QG methods in the sense that the question generation process is based on linguistic features (i.e., SRLs). However, instead of using rules to act on linguistic features, it uses the trained Seq2Seq models, which can model much more complex mapping relationships than human-generated rules. Since SRL-Seq2Seq is trained upon generalized sentence representations with linguistic labels instead of the original sentences, we conjecture that such trained model is more general and less domain-dependent than the Seq2Seq models trained on original QA sentences.

In this paper, we investigate whether this conjecture is true by training SRL-Seq2Seq on the SQuAD dataset (Du et al., 2017) and testing it on a dataset in a different domain. We compare SRL-Seq2Seq's performance with that of original Seq2Seq models in the same setting. In addition, since general-purpose large language models (such as GPT-4 (Achiam et al., 2023)) have shown remarkable success in various NLP tasks and domains. We will compare SRL-Seq2Seq's cross-domain generalization ability to that of GPT-4o.

2 SRL-Seq2Seq Model

The SRL-Seq2Seq model (Naeiji et al., 2023) combines SRL with Seq2Seq learning to enhance question generation. The process begins with the SRLer, which extracts predicate-argument structures (SRL labels) from training set answers. The Question2SRL mapper then aligns questions with these semantic representations, using either direct replacement (Hard-Question2SRL) or semantic sim-

Table 1: Evaluation results on **Car Manuals** with **BART** and **T5** as baselines (P, R and F mean Precision, Recall and F-score in %). SRL-Seq2Seq uses its Soft+C variation with BART or T5 as the Seq2Seq model. Better results in a pair-wise comparison between SRL-Seq2Seq and original Seq2Seq (BART or T5) are highlighted.

QG Method	BLEU-4			ROUGE-L			METEOR			BERT Score		
	F	P	R	F	P	R	F	P	R	F	P	R
BART	16.0	14.4	18.0	41.0	38.5	43.9	20.1	18.2	22.3	91.2	91.0	91.4
SRL-Seq2Seq with BART	**20.2**	**16.1**	**27.1**	**45.7**	**40.5**	**52.4**	**23.1**	**19.8**	**27.7**	**92.1**	**91.3**	**92.8**
T5	15.7	13.5	18.8	39.6	37.0	42.6	20.0	18.9	21.1	90.9	90.3	91.6
SRL-Seq2Seq with T5	**18.9**	**14.8**	**26.2**	**44.2**	**39.0**	**51.0**	**22.8**	**19.7**	**26.9**	**91.7**	**90.8**	**92.6**
GPT-4o	19.3	14.2	**30.3**	45.2	39.4	**53.2**	**24.2**	**21.6**	**27.7**	92.0	90.3	**93.7**

ilarity (Soft-Question2SRL). The Seq2Seq model is trained to convert SRL-labeled answers into corresponding question forms.

During inference, the SRLer extracts semantic representations from input sentences, which the Seq2Seq model converts into question representations. These are then translated back into natural language questions by the SRL2Question mapper. This method improves question generation for new, domain-specific datasets like the Car Manual Dataset, outperforming several baseline models. Section C illustrates this process with examples from (Naeiji et al., 2023) .

3 Datasets

We trained the model using the sentence-level SQuAD dataset (Du et al., 2017), where each answer is a single sentence. To evaluate the models, we employed the Car Manuals dataset (Naeiji et al., 2023). Further details about these datasets are provided in Section F.

4 Experiments and Findings

We conducted a cross-domain experiment in which models trained on the SQuAD dataset were utilized to generate questions from Car Manual. Specifically, we employed the BART-base and T5-small models from Huggingface's pre-trained models (Wolf et al., 2019), which contain 139 million and 60 million parameters, respectively.[1]. The details of the training can be found in Section E.

We use BLEU-4, ROUGE-L, METEOR and BERTScore as performance metrics. For each metric, we compute precision, recall and F-score, which are described in Appendix F. The results are shown in Table 1.

We would like to answer two questions in this study: (1) whether the generalization of training data using SRL in Seq2Seq learning makes the model more generalizable, and (2) how the results of SRL-Seq2Seq compares with the results from a much larger model (i.e., GPT-4).

To answer the first question, we compare the results from SRL-Seq2Seq with those of the Seq2Seq models without using SRL. Table 1 shows that SRL-Seq2Seq outperforms its corresponding baseline model in all metrics.[2] This indicates that incorporating SRL enhances the model's performance on out-of-domain datasets. Notably, the BART model, when combined with SRL, outperforms GPT-4o.

To compare with GPT-4, we employed GPT-4o in the zero-shot setting to generate questions from the Car Manuals test dataset. The prompt used in the experiment is shown in Appendix B, where we ask GPT-4o to generate questions that can be answered in an input sentence. As shown in Table 1, GPT-4o achieves the best recall in all metrics, but its precision is not as good (except with METEOR). Reviewing GPT-4's results, we found that questions it generated were more fluent than the ones generated by other models, but it had significant hallucination, with some questions lacking answers in the input text. GPT-4o often used its creativity to generate questions whose answers could not be found in the given input, which negatively impacted precision. However, GPT-4's ability to generate more questions increased the chances of matching ground truths, improving recall.

Looking at F-scores across the metrics, we found that SRL-Seq2Seq produces comparable results compared to GPT-4o. Considering SRL-Seq2Seq is a much smaller model than GPT-4o, the results of SRL-Seq2Seq are impressive.

[1]We use the smallest available model sizes of BART and T5 to avoid GPU memory errors.

[2]Note that the numbers reported in (Naeiji et al., 2023) are higher than the ones reported here. The reason is that they fine-tuned the models using the Car Manuals training set and then evaluated them on the Car Manuals test set. In contrast, in our investigation we fine-tune the models on SQuAD and then test the models on the Car Manuals test set. This cross-domain setup naturally leads to lower results due to the domain shift between training and testing data.

In summary, through the cross-domain experiment, we found that learning from SRL-labeled training data yielded a more generalized Seq2Seq model than training the Seq2Seq model using the original text and its performance was comparable to that of a much larger state-of-the-art large language model.

5 Limitations

Our investigation has some limitations. First, only one out-of-domain dataset is used in this investigation, which restricts the diversity and generalizability of our findings. Additionally, we employ the smallest versions of T5 and BART models due to GPU memory constraints. It would be interesting to see the results from larger models. Furthermore, the evaluation primarily relies on automatic metrics like BLEU and ROUGE, lacking human evaluation to fully assess qualitative factors such as fluency and coherence.

6 Acknowledgements

We would like to thank iNAGO Corporation for providing the Car Manuals dataset used in this research and for their collaboration on the research topic.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Yllias Chali and Sadid A Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics*, 41(1):1–20.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Twelfth International AAAI Conference on Web and Social Media*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. 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)*, pages 1342–1352.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. ChainCQG: Flow-aware conversational question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2061–2070, Online. Association for Computational Linguistics.

Payal Khullar, Konigari Rachna, Mukul Hase, and Manish Shrivastava. 2018. Automatic question generation using relative pronouns and adverbs. In *Proceedings of ACL 2018, Student Research Workshop*, pages 153–158.

Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018. Automating reading comprehension by generating question and answer pairs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 335–348. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7871. Association for Computational Linguistics.

Alireza Naeiji, Aijun An, Heidar Davoudi, Marjan Delpisheh, and Muath Alzghool. 2023. Question generation using sequence-to-sequence model with semantic role labels. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2830–2842.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequencepre-training. *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. 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*, pages 3901–3910.

A Generated Questions

An example of questions generated by different models on Car Manuals is shown in Table 2.

B GPT Prompt

This is the prompt that we used for generating questions by GPT:

"""Generate as many questions as you can from the following paragraph. Be careful that all the questions should be answerable by the paragraph. Do not generate questions that are not answerable by the paragraph. The questions should not be repetitive. {paragraph}"""

C Example of SRL-Seq2Seq Inference

Refer to Table 3.

D Dataset Details

The Car Manuals test set contains 1,869 QAs. For the SQuAD dataset the number of train, test and dev set is 70,484 QAs 11,877 QAs and 10,570 QAs respectively.

E Training Settings

During training on the SQuAD dataset, we optimized hyperparameters using a grid search approach as described by (Naeiji et al., 2023). The optimal learning rate (LR) and number of epochs (EP) were determined based on validation loss, with the best values being LR = 10^{-4} and EP = 10 for both T5 and BART models.

F Performance Metrics

To calculate precision, we evaluate each generated question against its corresponding reference questions, computing BLEU, METEOR, and ROUGE and BERTScore. For recall, we reverse the process: each reference question is compared with the corresponding generated question, and the scores are calculated accordingly. The F-score is the harmonic mean of precision and recall.

Model	Index	Questions
Input	-	Power is available whenever the vehicle is considered “ awake ” . the vehicle may be awake for many reasons such as when using features such as Summon , or when features such as Cabin Overheat Protection , Keep Climate On , etc . are on , as well as when the 12V battery is being charged or is in use , Leaving an accessory plugged in does not deplete the 12V battery
GPT-4o	1	when is power available in the vehicle ?
	2	what are some reasons the vehicle may be considered "awake" ?
	3	name a feature that can keep the vehicle awake.
	4	what happens when the 12v battery is being charged or is in use ?
	5	does leaving an accessory plugged in deplete the 12v battery ?
	6	what is one of the features mentioned that can keep the vehicle’s climate controlled ?
	7	what is the effect of cabin overheat protection on the vehicle’s state ?
	8	is the 12v battery affected by accessories being plugged in ?
	9	what does the term "awake" refer to in the context of the vehicle ?
T5	1	what is available whenever the vehicle is considered "awake" ?
	2	what is one reason the vehicle may be awake ?
	3	Leaving an accessory plugged in does not deplete what ?
BART	1	what is available whenever the vehicle is considered awake ?
	2	what is an example of a reason a vehicle may be awake ?
	3	what does leaving an accessory plugged in do ?
T5-SRL	1	what is an example of a feature that can be awakened ?
	2	when is power available ?
	3	what is an example of a feature that may be awakened ?
	4	what is one reason the vehicle may be awake ?
	5	what is an example of a feature that can be used to protect a vehicle ?
	6	what does leaving an accessory plugged in don’t deplete the 12v battery ?
BART SRL	1	what is an example of a reason a vehicle might be awake ?
	2	how does leaving an accessory plugged in deplete the 12v battery ?
	3	what is available whenever the vehicle is considered ‘ ‘ awake ” ?
	4	what is an example of a reason a the vehicle might be awake ?
	5	what does leaving an accessory plugged in do ?
	6	when is power available ?
	7	what is available whenever the vehicle is considered awake ?
Ground Truth	1	when can i charge my devices ?
	2	when can i use the usb ports ?
	3	my phone is dead , when can i charge my devices ?

Table 2: Ground truth and generated questions for one sample of Car Manual dataset

Table 3: Examples of sentence $\widehat{a}$, its semantic representation $\widehat{a}_{sem}$, the outcome $\widehat{q}_{sem}$ generated by Seq2Seq, the question $\widehat{q}$ converted from $\widehat{q}_{sem}$, and ground-truth question Q_t from the Car Manuals dataset.The table is taken from (Naeiji et al., 2023)

$\widehat{a}$:	before placing a child in the child restraint , make sure it is securely held in place .
$\widehat{a}_{sem}$:	before placing [ARG1] [ARG2] , make sure it is securely held in place .
$\widehat{q}_{sem}$:	what should i do before placing [ARG1] [ARG2] ?
$\widehat{q}$:	what should i do before placing a child in the child restraint ?
Q_t:	what should i do before placing a child in the child restraint ?