Vocabulary-constrained Question Generation with Rare Word Masking and Dual Attention

Emil Biju
ee17b071@smail.iitm.ac.in
Indian Institute of Technology Madras
Chennai, India

ABSTRACT

Question generation is the task of generating questions from a text passage that can be answered using information available in the passage. Known models for question generation are trained to predict words from a large, predefined vocabulary. However, a large vocabulary increases memory usage, training and inference times and a predefined vocabulary may not include context-specific words from the input passage. In this paper, we propose a neural question generation framework that generates semantically accurate and context-specific questions using a small-size vocabulary. We break the question generation task into two subtasks namely, generating the skeletal structure of a question using common words from the vocabulary and pointing to rare words from the input passage to complete the question.

ACM Reference Format:
Emil Biju. 2021. Vocabulary-constrained Question Generation with Rare Word Masking and Dual Attention. In *8th ACM IKDD CODS and 26th COMAD (CODS COMAD 2021), January 2–4, 2021, Bangalore, India.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3430984.3431074

1 INTRODUCTION

The task of question generation has gained prominence in recent years in the field of natural language generation due to its potential for application in automated tutoring and interviewing systems, chatbots, virtual assistants, etc. Traditional approaches have relied on heuristic rules based on syntactic parsing [2], part-of-speech tagging, semantic role labelling [1] and other structural patterns in the input passage. More recent approaches make use of LSTM/GRU-based neural encoder-decoder architectures that generate a question sequence by predicting words from a large vocabulary in successive timesteps. However, a large vocabulary increases the number of parameters, computational requirements and training/inference times of the network. Studies [4] show that 74-94% of conversations and written text in English are covered by the 1000 most common words. Therefore, we propose a two-step mechanism in which an encoder-decoder framework first generates the outline of a question by predicting words from a 1000-word vocabulary, followed by a recurrent neural network that copies words from the input passage to fill the gaps.

CODS COMAD 2021, January 2–4, 2021, Bangalore, India

ACM ISBN 978-1-4503-8817-7/21/01.
https://doi.org/10.1145/3430984.3431074

2 MODEL

For this work, we obtain question-answer pairs from the Stanford Question Answering Dataset (SQuAD) [3] to train our model. We first define the set of 1000 most commonly used English words as our vocabulary $\mathcal{V}$. We then prepare our target question sequences by masking all the remaining words with a <BLANK> tag. Sentences from each passage in the dataset are passed to the model as a sequence of ELMo word embeddings.

The **Question Structuring module** creates the skeletal structure of a question. It uses a Bidirectional LSTM encoder to encode the input sentence into a set of vectors, $\boldsymbol{H}$. The inputs to the decoder at each timestep are the embedding of the previously predicted word and the context vector obtained by applying Bahdanau attention over the encoder outputs. The decoder is a unidirectional LSTM that could either predict a common word from $\mathcal{V}$ or the <BLANK> tag at each timestep to generate the output sequence O.

The **Word Pointer module** replaces the <BLANK> tags in O with context-specific words from the input sentence. The outputs of the decoder LSTM in the previous module are passed to a Bidirectional LSTM to generate a set of encoded vectors, from which we define $\boldsymbol{Q} = \{q_i\}$ as the set of vectors such that O contains a <BLANK> tag at position i. Let $\boldsymbol{K} = \{k_j\}$ be the set of vectors from $\boldsymbol{H}$ corresponding to words $\{w_j\}$ from the input sentence that are not in $\mathcal{V}$. Using the Bahdanau attention mechanism, we compare each $q_i \in \boldsymbol{Q}$ with every $k_j \in \boldsymbol{K}$ and obtain attention scores $\alpha_{i,j}$. The <BLANK> tag at position i in O is replaced with the input sentence word w_{j^*} that received the highest attention score, i.e., $j^* = \arg\max_i \alpha_{i,j}$.

3 RESULTS AND CONCLUSION

On the test set,we observed an 8.8% higher BLEU score than an encoder-decoder LSTM framework with attention that uses the same vocabulary size and a nearly equal BLEU score when the LSTM framework uses 10 times our vocabulary size. Thus, we have demonstrated the efficacy of our network to generate context-specific questions using a size-constrained vocabulary. This technique can be extended to similar tasks like translation and summarisation.

REFERENCES

[1] Karen Mazidi and Rodney Nielsen. 2014. Linguistic considerations in automatic question generation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 321–326.
[2] Ruslan Mitkov et al. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing.* 17–22.
[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).
[4] Fred Joyce Schonell. 1956. *A Study of the Oral Vocabulary of Adults. An Investigation Into the Spoken Vocabulary of the Australian Worker.[By FJ Schonell and Others.].* Brisbane; University of London Press: London; printed in England.