End-to-end Question Generation to Assist Formative Assessment Design for Conceptual Knowledge Learning

Jinjin Zhao, Weijie Xu, Candace Thille

Abstract — **Formative assessment can be used by learning designers to evaluate a learners' comprehension, learning needs, and learning progress during a lesson, unit, or course. The general goal of a formative assessment is to collect detailed information that can be used to improve instruction and learning while learning is happening. Designing effective formative assessments for complex or technical knowledge can be difficult for a learning designer who does not have sufficient breadth or depth of expertise in the subject. The goal of this work is to provide assistance to designers in understanding the technical details of a subject and in constructing meaningful formative assessments. We propose an end-to-end solution that leverages text summarization, question generation, and context finetuning techniques to provide such assistance. In our solution, text summarization is applied to a text block to derive the main concept of the assessment. Question generation is then applied to both the text block and the main concept to generate a question. Human intervention is applied after the text summarization module to improve question quality. We use quantitative and qualitative measures to test various techniques in both the text summarization and question generation steps. The techniques include transformer-based solutions, sequence-to-sequence text generation, and contextualization of an NLP task. We demonstrate the solution with a use case from workforce learning. We also report findings on the effectiveness of the different approaches.**

Index Terms—**question generation, text summarization, formative assessment generation, natural language understanding.**

I. Introduction

Formative assessment refers to the diagnostic use of assessment to provide feedback to teachers and learners over the course of instruction [11]. When teachers know how learners are progressing and where they are having trouble, they can make necessary instructional adjustments, such as reteaching, trying alternative instructional approaches, or offering more opportunities for practice. Feedback given directly to learners as part of formative assessment helps them become aware of any gaps that exist between their desired goal and their current knowledge and guides them through actions necessary to obtain the goal [12]. Designers apply learning theories, design practices, along with the subject matter knowledge, to construct formative assessments. When designers do not have the domain specific knowledge, it is difficult for them to understand the subject matter and to deliver effective formative assessment. In this work, we aim to leverage natural language processing (NLP) tools to help designers deliver quality assessments. We focus on generating useful questions given a piece of content with two steps, 1. summarizing a piece of content, and 2. generating meaningful questions that can be altered to fit the learning and assessment needs.

NLP techniques developed for tasks such as text summarization and question generation have shown promising results in analyzing the text and deriving insights. We are interested in examining various NLP techniques in the context of formative assessment creation for conceptual knowledge learning. The techniques we examine include state-of-the-art transformer-based summarization techniques (both extractive and abstractive), sequence-to-sequence text generation, and language model contextualization. For a text summarization task, the challenge is to understand the semantics and compress the content. As mentioned in the work [1], statistical features or graphical relationships between key phrases is insufficient in understanding and representing semantics when the text is free-form and less well structured, so we delve into different methods within the transformer-based language model (LM) solution in this work. There are extractive and abstractive methods to compress the content. Extractive methods summarize a given piece of content by identifying the key phrases and ranking them with certain measures. Extractive methods may either over extract or under extract the key concepts if the 'similarity' thresholding is not optimized in a specific use case. Abstractive methods work by generative modeling and rephrasing the concepts after understanding them. Abigail, Liu, and Manning [6] proposed using a pointer-generator network to produce a generative summary of a given piece of content. The shortcoming of that method is that the summarization quality is mostly determined by how the LM is pre-trained and fine-tuned. In other words, there is less quality control over the resulting summary. In this work, we test both extractive and abstractive LM-based summarization methods and examine their effectiveness in the context of formative assessment creation.

In recent years, sequence-to-sequence techniques are emerging for text generation. A variety of neural network architectures have been proposed to tackle various tasks. Sutskever [8] proposed a sequence-to-sequence framework to extract the relationship between sentence and question pairs. Zhao [10] employed a maxout pointer mechanism with gated self-attention encoder to derive the pattern of generating a question from a given content and answer pair. Fine-tuning a LM for a downstream task with a sequence-to-sequence framework is well applied. The fine-tuning step learns the

Jinjin Zhao is a Senior Machine Learning Scientist at Amazon (email: jinjzhao@amazon.com)

Weijie Xu is a Machine Learning Scientist at Amazon (email: weijiexu@amazon.com)

Candace Thille is Director of Global Learning and Development organization at Amazon (cthille@amazon.com)

task pattern from its input and output in a supervised manner. In this work, we test different fine-tuned LMs (embedding finetuning and task finetuning) for generating questions. We also test whether or not fine-tuning answer format can improve question quality in a given context.

The research questions we seek to answer are:

1. Does text summarization produce a meaningful and succinct concept for conceptual knowledge? Which work better, extractive or abstractive methods?
2. Is question generation output helpful questions in assisting formative assessment design?
3. Is finetuning an answer format helpful in producing a quality question?
4. Which steps are most critical in the end-to-end solution: text summarization, question generation, or human intervention in between?

II. Approach

We propose an NLP pipeline that has four components to generate questions given a piece of content. The first component is text summarization. It aims to derive the summary of concepts for a given piece of content. The second component is question generation. It produces questions given the content and the summary derived in the first component. The third component is human intervention between summarization and question generation. Designers are encouraged to understand the summary and paraphrase it into a summary. The paraphrased summary is fed into the question generation task along with the content. The fourth step is to finetune the LM, including the embedding representation and the NLP task format, for better summary and generated question quality.

A. Concept Summarization

The goal of summarization is to provide a shorter version of the content as a proxy for the 'answer' that is required in question generation task. We define an acceptable summarization as either 1. a succinct summary highlighting the main idea, or 2. key concepts of the main idea. In industrial applications, a text pair in the format of <paragraph, summary> is usually unavailable and expensive to prepare. In other words, a training process for a specific summarization task is expensive to conduct. Therefore, we leverage the existing LMs that are trained on public datasets with text pair <paragraph, summary> data and perform an inference step to generate the summary. We apply both abstractive and extractive methods to summarize a given piece of content. In the abstractive category, we use T5 [5] and BART [3] based LMs. Both models use a transformer-based architecture where the attention [9] mechanism is adopted. Distillbart [7] is a distilled BART model that has a smaller size and is proposed for practical use. We tested top performers from huggingface.co, including

't5-small(lm-summ-1)',
'sshleifer/distilbart-cnn-12-6(lm-summ-2)',
'sshleifer/distilbart-xsum-12-6(lm-summ-3)',
'facebook/bart-large-cnn(lm-summ-4)'.

In the extractive category, based on our prior work findings [1], semantically contextualized LM has the highest accuracy and the most robust performance, we use a BERT based key phrase extractor. We hypothesize that the most semantically similar phrases (compared to paragraph semantics) are the key phrases of the given paragraph. Thus, we first extract potential key phrases from the paragraph with segmentation techniques (tokenization, stemmation, lemmati-zation, etc.), and then compare each phrase embedding with paragraph-level embedding. At the end, we apply an adversity-based maximum marginal relevance ranking algorithm to remove duplicates.

B. Question Generation with LM fine-tuning

Given the 'answer'(summarization) and the content, the question generation module aims to produce a question that can be used to assess the knowledge state of a learner. We use text generation technique and text-to-text fine-tuning practice to produce the question. Due to the lack of the <content, answer | question> format data, we leverage existing LMs that are trained on a public dataset with question generation tasks and conduct an inference step to generate the question. We apply transformer-based architectures and LMs to represent the text and the language task. We use the same LM T5 as in the summarization. The LM is fine-tuned on SQuADv1 dataset [13] with highlighted answer format (e.g., <hl> 42 <hl> is the answer to life, the universe and everything. </s>). We test different fine-tuned LMs, 'valhalla/t5-base-qg-hl'(lm-qg-1),
'valhalla/t5-small-qg-hl'(lm-qg-2).

In SQuADv1 dataset, the answer is a span from the original paragraph and the answer format is predefined. In our context, the answer might be a paraphrase or a generative summary of the original text. The answer format is prepared differently from the public dataset (e.g., 42 is a meaning number <sep> 42 is the answer to life, the universe and everything). Thus, we fine-tune the LM model with its desired answer format and see if the format fine-tuning step makes any difference on the final question generation.

C. Human Intervention in between

Human intervention is encouraged after the summarization step and before the question generation step. During intervention, designers can learn about the concept from the generated summary, integrate the information into their own knowledge, and rephrase the summary in a more natural way. In the question generation step, both the content and rephrased answer are used to generate the question. We test whether a paraphrasing step contributes to a better generated question.

D. Measures

For quantitative measures, we use Bilingual Evaluation Understudy Score (BLEU Score) [4] and a Metric for Evaluation of Translation with Explicit ORdering (METEOR Score) [2] to quantify the quality of the summarized answers and generated questions. BLEU score is a metric for evaluating how similar a generated sentence is to a reference sentence. A perfect match results in a score of 1.0, and a perfect mismatch results in a score of 0.0. It is quick and inexpensive to calculate. It's highly correlated with human evaluation and has been widely adopted. In our application, the generated summary and questions are short in length, we thus only consider unigram and bigram in our analysis. METEOR score was proposed to measure machine translation hypotheses by aligning the sentence to one or

more reference translations. Alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases. Segment and system level metric scores are calculated based on the alignment between hypothesis-reference pairs. For qualitative measures, we leverage human evaluation to rate the generated question quality. We define three levels (score 1 to 3) of quality: Score 1: not useful; Score 2: a bit useful; Score 3: very useful. Two to three raters (subject matter experts) are assigned to rate the questions. If there is disagreement on the scoring, raters re-evaluate until they reach to consensus.

III. EXPERIMENTS

A. Case Study

Application context: We apply the proposed solution to two courses that are designed to introduce AWS infrastructure to software developers. Since the concepts are technical and targeted for software developers, it is difficult for course designers (who has no background knowledge or skills on software development) to quickly understand the content and design formative assessments. We focus on content that is about AWS infrastructure concept introduction in breadth and depth. The procedural knowledge about how to get hands-on to apply the concepts and build products is out of scope for this research. For each course, we have ten paragraphs that introduce related concepts. The average paragraph length is 82 words. The shortest paragraph has 56 words. The longest paragraph has 103 words. Each paragraph is focused on one particular concept and its sub concepts or related knowledge.

Experiment setup: In extractive summarization, we apply BERT based key phrase extractor and test Top N (N=1,3,5) key phrases to represent the key concept and its sub concepts. In abstractive summarization, we apply different LMs that are fine-tuned on text summarization tasks. LMs include lm-summ-1 to lm-summ-4 that are introduced in section Approach A. We apply different LMs that are fine-tuned on question generation tasks. LMs include lm-qg-1 and lm-qg-2 as introduced in section Approach B. We also further fine-tune LMs with targeted pair format with dataset SQuADv2. We use the first 2000 entries of the dataset for fine-tuning, with epoches=20 and batchsize=1.

We use one example (paragraph) to demonstrate the process of generating questions. The paragraph, as shown in Table 1, introduces Amazon AWS VPC service on: 1. what it is; 2. what advantage it has; 3. what other benefits it brings. We expect to generate questions around these three aspects. Table 1 illustrates its content, generated answers, and the generated questions with each method. As shown, extractive summarization methods deliver valid questions on the first aspect - what VPC is. There is a slight difference in phrasing the question based on the Top N (1,3,5) hyperparameter. Abstractive summarization methods generate answers in different lengths which are determined by how the selected LM is trained. The generated questions cover two aspects: 1. what VPC is and 2. what other benefits the AWS VPC brings. Table 1 detail is as follows, M*: Methods. Methods 1: extractive (Top 1); 2: extractive (Top 3); 3: extractive (Top 5); 4: abstractive(lm-summ-1); 5: abstractive (lm-summ-2); 6: abstractive (lm-summ-3); 7: abstractive (lm-summ-4). All with(lm-qg-1). 8* is from human.

B. Quantitative Measures

For quantitative measure, we ask two subject matter experts to write their answers and questions as reference. We use the following text for summaries and questions as reference. References for summaries are 1. virtual networking environment 2. AWS VPC benefits. References for questions are 1. What is AWS VPC? 2. How can Amazon VPC help you? As shown in Table 2, from both BLEU and METEOR measures, the extractive method with Top 1 performs the best, for both answers and questions. We also observe that the METEOR score has a higher correlation between generated text and human prepared references than the BLEU score. Since the BLEU score relies on a precision metric, the longer the answer is compared to the reference, the smaller the score would be. The METEOR score focuses more on recall metric. As a result, the METEOR score would still stay high if the keyword appears in the generated text, while the BLEU score would decrease if the generated text covers more irrelevant information.

TABLE 1: PARAGRAPH, ANSWERS, AND QUESTIONS.

Paragragh		
Amazon VPC allows you to provision a logically isolated section of the Amazon Web Services (AWS) cloud, where you can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking environment including selection of your own IP address ranges, creation of subnets, and configuration of route tables and network gateways. You can also create a hardware Virtual Private Network (VPN) connection between your corporate datacenter and your VPC and leverage the AWS cloud as an extension of your corporate datacenter.		
M*	**Answers**	**Questions**
1	virtual networking environment	What is the AWS VPC?
2	virtual networking environment, network gateways, route tables	What does Amazon VPC do for you?
3	virtual networking environment, network gateways, route tables, subnets, corporate datacenter	What does Amazon VPC allow you to provision?
4	You can create a virtual private network (VPN) connection between your corporate datacenter and your VPC. Amazon VPC allows you to provide a logically isolated section of the Amazon Web Services (AWS) cloud. You have complete control over your virtual networking environment.	How can you leverage the AWS cloud as an extension of your corporate datacenter?
5	Amazon VPC allows you to provision a logically isolated section of the Amazon Web Services (AWS) cloud. You can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking environment.	What does Amazon VPC allow you to provision?
6	Amazon VPC is a virtual network that allows you to connect to your corporate datacenter with the cloud.	What does Amazon VPC do for you?
7	Amazon VPC allows you to provision a logically isolated section of the Amazon Web Services (AWS) cloud. You have complete control over your virtual networking environment. You can also create a hardware Virtual Private Network (VPN) connection between your corporate datacenter and your VPC.	How can you leverage the AWS cloud as an extension of your corporate datacenter?
8*	AWS VPC benefits	What is AWS VPC?

C. Qualitative Measure Task

We report the human grading result of the generated questions in Table 3. Each row is for a single paragraph. Each column is for a different method for that paragraph. As shown in Table 3, the average grading across 20 paragraphs shows that both abstractive and extractive summarization achieve good performance (2.6 out of 3). For extractive methods, the final question quality is insensitive to the Top N hyperparameter. For abstractive methods, lm-summ-3 and lm-summ-4 achieve better results. The difference among \#4-\#7 and \#8-\#11 shows that both LM can produce quality questions and lm-qg-2 performs better. There is no significant difference between \#8-\#9 and \#10-\#11, which indicates that the fine-tuning step on answer formatting does not contribute to the question quality. Table 3 reports 11 methods for 20 paragraphs. Row: Paragraph Index (P). Column: Method Index (M). Row A: averaged performance for a method across 20 paragraphs. M#1: extractive (Top 1);
M#2: extractive (Top 3);
M#3: extractive (Top 5);

M#4: abstractive (lm-summ-1);
M#5: abstractive (lm-summ-2);
M#6: abstractive (lm-summ-3);
M#7: abstractive (lm-summ-4);
M#8: abstractive (lm-summ-3);
M#9: abstractive (lm-summ-4);
M#10: abstractive (lm-summ-3);
M#11: abstractive (lm-summ-4);
M#1-7, lm-qg-1. M#8-11, lm-qg-2.
M#10-11, answer format fine-tuned.

TABLE 2: BLEU SCORE AND METEOR SCORE

#	bs @a	bs @q	ms @a	ms @q
1	1.0	0.8634	0.9814	0.9146
2	0.7178	0.7002	0.8660	0.4189
3	0.5366	0.5943	0.7957	0.4122
4	0.4521	0.2885	0.2977	0.3758
5	0.2932	0.5943	0.6722	0.4122
6	0.5020	0.7002	0.4481	0.4189
7	0.4041	0.2885	0.1401	0.3758

TABLE 3: QUALITATIVE EVALUATION (HUMAN RATING)

#	1	2	3	4	5	6	7	8	9	10	11
1	2	2	2	3	3	2	3	3	3	3	3
2	3	3	3	3	3	3	3	3	2	3	2
3	3	3	3	2	3	3	3	3	3	3	3
4	1	2	2	2	3	2	2	3	2	3	2
5	3	2	2	2	2	2	2	3	3	3	3
6	2	2	2	2	2	3	3	3	3	3	3
7	1	1	2	2	2	3	3	2	3	2	3
8	3	3	3	2	2	3	3	3	3	3	3
9	2	2	2	2	2	2	2	2	2	2	2
10	3	3	3	2	3	3	2	3	3	3	3
11	2	2	2	2	3	3	3	3	3	3	3
12	3	3	3	3	3	3	3	3	3	3	3
13	2	2	2	2	2	2	3	3	3	3	3
14	2	2	2	2	2	3	2	3	2	3	2
15	2	3	3	2	2	3	2	3	3	3	3
16	3	3	3	3	2	2	3	2	3	2	3
17	3	3	3	2	2	3	3	3	3	3	3
18	3	3	3	4	3	3	3	3	3	3	3
19	2	3	3	2	2	2	2	2	3	2	3
20	2	2	2	2	2	2	2	3	2	3	2
A	2.4	2.5	2.5	2.3	2.4	2.6	2.6	2.8	2.8	2.8	2.8

IV. CONCLUSIONS AND FUTURE WORK

Regarding the research questions, we find that both extractive and abstractive methods are able to generate a meaningful summary. Extractive methods are able to extract key concepts as the answer and abstractive methods are able to paraphrase and depict the concept in a more natural way with more details around the concepts. The question generation step is able to produce meaningful examples with different expressions. The generated questions carry the same semantics and are able to offer help to designers in creating formative assessments. Another finding around question generation is that regardless of how the answer is expressed in the summarization step, question generation produces semantically similar expressions around the main concepts. This indicates that the details around the concepts (as part of the answer) are not contributing to the question generation result. In other words, the question generation step is robust to the answer preparation as long as the key concept is captured at a high level. We also find that fine-tuning the answer format does not bring additional value to the question quality. The question generation step is robust irrespective of how the answer and content are fed into the algorithm, whether using highlighted format (<hl>) or other separating symbols (<sep>). It seems that the question quality is less sensitive to the answer preparation step compared to the question generation step. More experiments are needed to test different question generation algorithms. Human intervention between answer preparation and final question generation does not show significant help in our studies. Designers found that the summarized answer was useful for them to get some idea of what the technical content is about. They also thought that the details of the key concepts were not contributing as much assistance as the key concepts when they prepared the formative assessments. This work is a starting point of applying SOTA NLP tools to assist designers in creating formative assessments for conceptual knowledge learning, especially when the subject matter is technical and complex. To provide assistance to designers in these cases, we proposed an end-to-end question generation framework, from understanding the content to generating a summary to generating questions. We examined how effective each step of the framework is in producing quality questions for two courses implemented in a corporate upskilling setting. We plan to continue working on applying this approach to other conceptual knowledge learning or procedural knowledge learning to explore its limitation and improve it. Representation learning for the desired knowledge is required for the algorithm to understand the subject matter. For example, we need to either fine-tune the LM or learn the representation (for instance, medical application and software coding application require representation learning of the domain specific code) before we can leverage the existing NLP tools.

REFERENCES

[1] Zhao, Jinjin, Kim Larson, Weijie Xu, Neelesh Gattani, and Candace Thille. "Targeted Feedback Generation for Constructed-Response Questions." (2020)

[2] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65-72. 2005

[3] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019)

[4] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318. 2002

[5] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019)

[6] Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." *arXiv preprint arXiv:1704.04368* (2017)

[7] Shleifer, Sam, and Alexander M. Rush. "Pre-trained summarization distillation." *arXiv preprint arXiv:2010.13002* (2020).

[8] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." In *Advances in neural information processing systems*, pp. 3104-3112. 2014

[9] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017

[10] Zhao, Yao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. "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.

[11] Boston, Carol. "The concept of formative assessment." *Practical Assessment, Research, and Evaluation* 8, no. 1 (2002): 9

[12] Black, Paul, and Dylan Wiliam. "Developing the theory of formative assessment." Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education) 21, no. 1 (2009): 5-31.

[13] Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint arXiv:1606.05250* (2016).