Multimodal Question Generation using Multimodal Adaptation Gate (MAG) and BERT-based Model

1st Muhammad Farhan Akbar
School of Computing
Telkom University
Bandung, Indonesia
farhanakbar@student.telkomuniversity.ac.id

2nd Said Al Faraby
School of Computing
Telkom University
Bandung, Indonesia
saidalfaraby@telkomuniversity.ac.id

3rd Ade Romadhony
School of Computing
Telkom University
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

4th Adiwijaya Adiwijaya
School of Computing
Telkom University
Bandung, Indonesia
adiwijaya@telkomuniversity.ac.id

***Abstract*—Question Generation (QG) is a task to generate questions based on an input context. Question Generation can be solved in several ways, ranging from conventional rule-based systems to recently emerging sequence-to-sequence approaches. The limitation of most QG systems is its limitation on input form, which is mainly only on text data. On the other hand, Multimodal QG covers several different inputs such as: text, image, table, video, or even acoustics. In this paper, we present our proposed method to handle the Multimodal Question Generation task using an attachment to a BERT-based model called Multimodal Adaptation Gate (MAG). The results show that using the proposed method, this development succeeds to do a Multimodal Question Generation task. The generated questions give 16.05 BLEU 4 and 28.27 ROUGE-L scores, accompanied by the human evaluation to judge the generated questions from the model, resulting in 55% fluency and 53% relevance.**

***Index Terms*—Question Generation, Multimodal Question Generation, BERT**

I. Introduction

Question Generation (QG) is a part of Natural Language Processing (NLP) where it aims to automatically generate questions from natural language input like text, e.g. a sentence or a full paragraph, and its target answer [1]. Question Generation received considerable interest in recent years, especially in industrial and academic fields. A study by Chan et al. shows that using a pre-trained model of BERT (Bidirectional Encoder Representation from Transformer) can solve the Question Generation problem. They build a model with a BLEU score of 21.04 on the SQuAD dataset [2]. Currently, the best result of Question Generation on the SQuAD dataset is informed in [3], which has a BLEU-4 score of 25.41.

Researches on Question Generation rely on text-only input and do not consider other forms of input such as images or tables [4]. In some cases, another form of input like images are needed to understand the generated questions and the possible answer to answer those questions [5]. For example, the reader will find it difficult to answer the question from a passage "What is the name of a famous painting where there is a woman with a sad smile?" without its context image. Multimodal Question Generation is a solution to this problem where the context is in many forms. Multimodal Question Generation refers to the task that automatically generates questions from multimodal input, where multimodal means that having more than one form, can be modalities from text, vision, or acoustic [6].

Recently, BERT succeed to solve multiple NLP problems. In [6], they succeed to build an attachment to BERT for multimodal language, an area in NLP that centered on modeling face-to-face communication. This attachment is called Multimodal Adaptation Gate (MAG), which allows BERT to accept not only text input, but also other forms of input like images or sounds during fine-tuning. A layer of the BERT network that had MAG added to it is known as MAG-BERT, which forms a representation of multimodal information with corresponding lexical vectors. However, in the study by Rahman et al., MAG-BERT is used for the sentiment analysis task, not a generation task.

A study on QG using the BERT model has been conducted by [7]. There are three architectures proposed to tackle the QG task. The result shows that they succeed and outperform the previous approach of RNN-based regarding automatic evaluation metrics like BLEU, ROUGE, and METEOR. Even though the model proposed is simple, it achieves the best performance on sentence-level and paragraph-level input.

We proposed a solution to overcome problems with multimodal inputs on Question Generation from the MMQA (MULTIMODALQA) dataset, which contains figures and text. MMQA provides a large-scale QA dataset that requires integrating information from text, and images, where 35.7% of the entire dataset requires cross-modality reasoning [8]. We proposed an attachment of MAG from [6] to the BERT-based model for the Question Generation task from [7]. Generated questions from the proposed method were evaluated using automatic evaluation and human evaluation to ensure the quality of the generated questions. The contributions of this study include:

- Perform the Multimodal Question Generation on text and image modalities from the MMQA dataset using the proposed model which adds an attachment of MAG to the BERT-based model
- Conduct a human evaluation to evaluate the result of the generated questions from the proposed model

II. Related Works

The Question Generation task is typically approached as a sequence-to-sequence learning problem that involves converting a sentence from a passage of text into a question [9]. As an approach to the Question Generation problem from [2], Chan et al. proposed a model based on BERT, which is structured from [10], generating more logical and fluid semantics. But, those studies only consider single modality context from text passages, and their context answer, making limitations for inputs that are more than single modality or multimodal.

To fuse multimodal inputs, every modality requires to be converted into a learnable tokenizer, then cross-modality understanding can be achieved [11]. Another approach in [6], the proposed framework successfully demonstrates a multimodal sentiment analysis task where, in the process, multimodal input can be accepted to a pre-trained BERT and XLNet model by attaching an attachment called Multimodal Adaptation Gate (MAG), which allows those models to accept more than one modality.

Based on all aforementioned studies on QG, no approach has been proposed to process multimodal inputs (text and image) to generate a question.

A. Multimodal Adaptation Gate (MAG)

In multimodal language, words are accompanied by visual and acoustic information, such as gestures and prosody. These nonverbal behaviors can alter the meaning of words and how they are understood within the context of a conversation, which captures latent concepts for individual words. The position of a word in this space is based on the word's meaning in a linguistic structure, like a sentence. The Multimodal Adaptation Gate (MAG) takes into account the impact of nonverbal behaviors on a word's position in the semantic space with a displacement vector. This vector reflects the trajectory and magnitude of the changes in meaning or interpretation that are caused by nonverbal behaviors. [6].

MAG illustration in Figure 1 shows that MAG takes three inputs, lexical, visual, and acoustic. Let Z_i, A_i, and V_i denote those inputs, which are combined into bimodal factors $[Z_i; A_i]$ and $[Z_i; V_i]$ using vector concatenation. After obtaining lexical vectors that incorporate both acoustic and visual information, these are used to generate two gating vectors g_i^v and g_i^a.

$$g_i^v = R(W_{gv}[Z_i; V_i] + b_v) \tag{1}$$

$$g_i^a = R(W_{ga}[Z_i; A_i] + b_a) \tag{2}$$

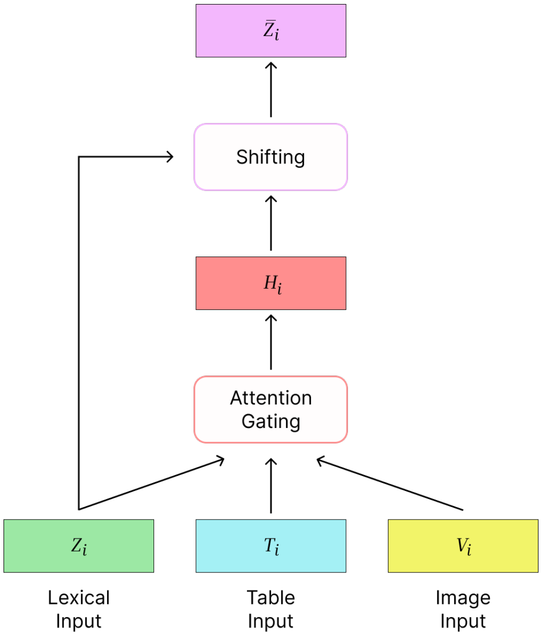

Fig. 1. Multimodal Adaptation Gate (MAG)

W_{gv} and W_{ga} respectively are visual weight matrices and acoustic weight matrices, while b_v and b_a are scalar biases. $R(x)$ is a non-linear activation function. The two gating vectors are designed to convey important information from the visual and acoustic modalities that are influenced by the lexical vector. Following this, a non-verbal vector H_i is created by combining A_i and V_i after they have been multiplied by their corresponding gating vectors.

$$H_i = g_i^a \cdot (W_a A_i) + g_i^v \cdot (W_v V_i) + b_H \tag{3}$$

W_a and W_v are weight matrices that represent the importance of acoustic and visual information, respectively, and b_H is a bias vector. The nonverbal vector H_i is then added to the lexical vector Z_i to produce a multimodal vector $\bar{Z}_i$.

$$\bar{Z}_i = Z_i + \alpha H_i \tag{4}$$

$$\alpha = min(\frac{||Z_i||_2}{||H_i||_2}\beta, 1) \tag{5}$$

β is a hyper-parameter that is determined through the cross-validation process. The L_2 norm of the vectors Z_i and H_i is denoted by $||Z_i||2$ and $||Zi||_2$, respectively. The scaling factor a is used to ensure that the impact of the nonverbal shift H_i remains within the desired range. At last, the normalization layer and dropout layer were applied to $\bar{Z}_i$.

B. BERT Overview

BERT (Bidirectional Encoder Representation from Transformer), a language representation,BERT (Bidirectional Encoder Representation from Transformer), is a language representation, focused on bidirectional representations from unlabeled text using a pre-training approach. BERT has two