CHAPTER 1

INDTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Objective and Hypothesis; (6) Assumption; (7) Scope and Delimitation; and (8) Importance of the study.

1.1 Rationale

Template extraction is a form of Information Extraction (IE) that aims to obtain patterns from data. A few studies considered that template extraction is beneficial in some tasks such as question template extraction which improves the performance of the question-answering model [1, 5]. Another advantage of question template extraction also pointed in [9] which stated that retrieving the template of a question sentence enhanced the ability of a question generation model to paraphrase a question sentence into several questions with the same meaning but different expressions. Based on those studies, it is implied that question template extraction is beneficial in solving various Natural Language Processing (NLP) problems. Several studies have been carried out on IE tasks that are related to template extraction. For instance, [2] shows that extracting key phrases from a document is possible with a sequence labeling approach. In addition, [9] also stated that the sequence labeling approach is capable of extracting a template from a question by utilizing Named Entity Recognition (NER) with Input-Output-Beginning (IOB) tagging scheme using Bidirectional Long Short-Term Memory (BiLSTM). These studies suggested that sequence labeling is beneficial in IE and could be implemented in template extraction.

1.2 Theoretical Framework

The advantage of question template extraction was also pointed out in [9] which stated that retrieving the template of a question sentence enhanced the ability of a question generation model to paraphrase a question sentence into several questions with the same meaning but different expressions. Question template extraction is suggested as an approach to obtain a generalization of a question sentence structure, hence the template could be utilized in multiple domains. Although IE in extracting question templates is less explored, several studies have been carried out on IE tasks related to template extraction. For instance, [2] shows that extracting key phrases from a document is possible with a sequence labeling approach. In addition, [9] also stated that the sequence labeling approach is capable of extracting a template from a question by utilizing Named Entity Recognition (NER) with Input-Output-Beginning (IOB) tagging scheme using Bidirectional Long Short-Term Memory (BiLSTM). These studies suggested that sequence labeling is beneficial in IE and could be implemented in template extraction.

1.3 Conceptual Framework/Paradigm

Based on the explanation from the previous section, the performance of the best model from those studies has not been tested in chunking question sentences. Hence, further exploration and analysis of question template extraction using text chunking as the sequence labeling approach is required. Based on limitations from prior studies related to text chunking, this study aims to propose a question template extraction model using a sequence labeling approach that has better performance than [22] in terms of ROUGE score. This study used data that are collected from the research that Cao and Wang conducted in 2021 related to open-ended questions [5]. This data consists of numerous question sentences compiled from multiple sources [5]. To extract question templates from the dataset, this study implemented a BiLSTM model as prior studies associated with sequence labeling suggested that Bidirectional Long Short-Term (BiLSTM) was proven capable of extracting text patterns with high accuracy [3, 11].

1.4 Statement of the Problem

Based on prior studies, a detailed explanation of the utilization of the sequence labeling approach in question template extraction tasks is not justified yet. The discussion in [5] only focused on the ability of question template extraction in diversifying generated questions. Furthermore, [9] only mentioned the general steps to extract question templates and only focused on using NER in the template extraction process, other tagging schemes such as text chunking have not been explored yet in those studies. In contrast with Part-of-Speech Tagging (POS Tagging) which categorizes each word, text chunking aims to categorize a group of words in a sequence to a label. Although, studies regarding text chunking as a sequence labeling approach have been carried out in the past years. In 2020, Wang investigated how to identify noun phrases in a text using Recurrent Neural Networks [22]. This study used English Treebank as the dataset and their best model achieved 93% in F-score.

1.5 Objective and Hypothesis

Based on those problems, this study aims to propose a novel question template extraction model using a sequence labeling approach that has a better performance compared to the current state-of-the-art model from [22] in Open Ended Question Dataset [5]. The evaluation was carried out to measure the performance of the proposed method by calculating Recall-Oriented Understudy for Gisting Evaluation (ROUGE).

The hypothesis of this study is that the proposed sequence labeling method in this study will have a higher performance than state-of-the-art models based on ROUGE score in Open Ended Question Dataset [5] in extracting template from a question sentence compared to [22], where chunking that implemented in their study was limited to noun phrase, while in this study, additional phrases such as verb phrase and adjective phrase also involved. In addition, [22] has not explored the CRF layer in their model considering that it played a vital role in the model's performance improvement as stated in prior sequence labeling studies [3], [11].

1.6 Assumption

Based on results from several studies conducted in the past, the proposed method is assumed to be superior to the current state-of-the-art model in extracting question templates using a sequence labeling approach since the proposed model in this study was trained using a dataset that involved more phrases than the dataset that was used in the prior study.

1.7 Scope and Delimitation

This study focused on solving question template extraction problems specifically using the sequence labeling approach. The dataset that was implemented in this study was limited to an open-ended question dataset which contained various question sentences. The phrases that were labeled in this study were limited to noun phrases, verb phrases, and adjective phrases. In addition, the model that was implemented in this study was limited to BiLSTM.

1.8 Significance of the Study

This study conveys a detailed explanation regarding template extraction specifically in question-related tasks. Also, it provides a novel question template extraction model superior to the current state-of-the-art model. By investigating how the template of a sentence could be extracted, this study could be a guide in constructing a question template extraction model. Additionally, the findings of this study could be a reference to determine the value of the hyperparameter of the model so that the performance would be optimal.