The Process Overview of Thai QA Framework

In this section, we describe a process of Thai question answering framework. The details of framework are explained and illustrated in Figure 1.

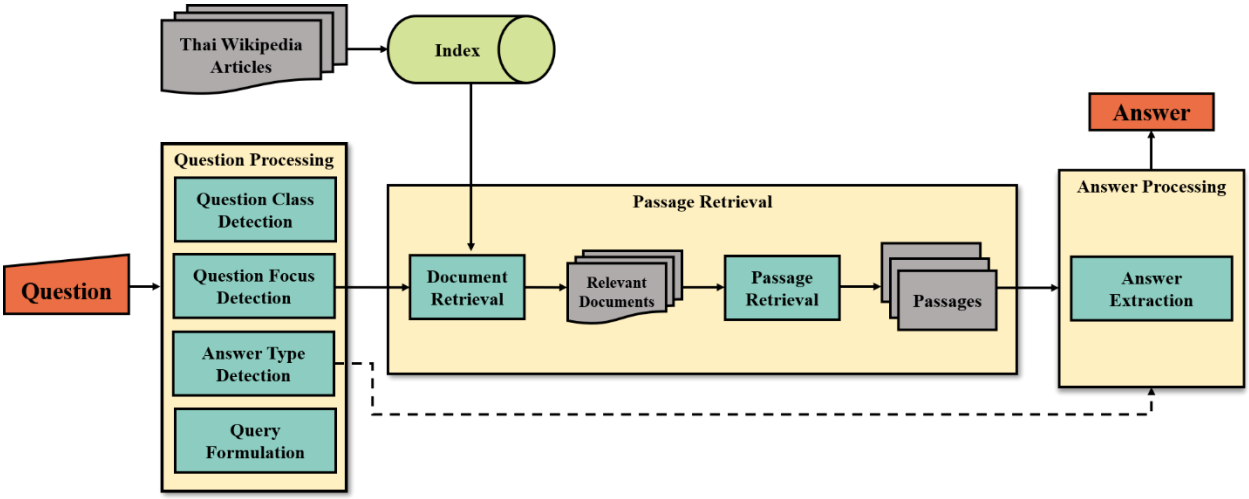

Figure 1. Thai Question Answering Framework

The key objective of Thai question answering framework is to automatically generate a correct answer of a question by finding relevant documents containing answer strings on Thai Wikipedia articles. Our proposed framework consists of three main components: (1) question processing (2) passage retrieval, and (3) answer processing. Each of the components is explained below.

1) Question processing component

The goal of this component is to analysis a question by detecting question type, answer type, and question focus. The question then is reformulated as queries before sending to the passage retrieval component to match relevant documents. This component is divided into four main modules: (1) question class detection (2) answer type detection (3) question focus detection, and (4) query formulation. The first module is the task of defining a class of the question such as a factoid question, a math question, and a yes/no question. The second module, the task of detecting a named entity type (person, location, datetime, etc.) of the answer using a rule-based technique with question words. For example, the word “Where” is quite consisting of matching locations. The third module applies a pattern matching approach to extract a focus word in a question that are likely to be replaced by the answer. For the final module, the module aims to transform a question into queries to send to the passage retrieval component. The question is normalized by tokenizing and removing stop words to be a suitable form for increasing the accuracy of finding a set of text passages on relevant documents.

2) **Passage retrieval component**

This component aims to retrieve a set of text passages that might contain answer strings from the Thai Wikipedia articles. The transformed queries are passed to this component based on an okapi BM25 technique to perform the document and passage retrieval modules. Then, the set of candidate text passages will be selected in the 5 highest average scores and ranked by a sentence similarity algorithm.

3) **Answer processing component**

The final component of our framework is to extract a correct answer on the set of candidate text passages. We apply a named entity tagger and an extraction pattern approach to parse and identify the expected answer types on the candidate text passages. Finally, the component returns the best correct answer matching with the named entity type from the answer type detection module.

The Evaluation Results

To evaluate the performance of our proposed framework, we prepared the evaluation dataset of 5,000 question-answer pairs created by human annotators from Thai Wikipedia articles. The experimental setting was conducted for two tasks: (1) document retrieval task and (2) document reader task. For the task of document retrieval, we used F-measure that is commonly used to evaluate the performance in information retrieval. From the experimental result shows that our technique of the proposed framework achieved 71.24% with 1-best accuracy. For the task of the document reader, we use the exact match (EM) metric that computes on common substring of word level between the predicted answers and the gold answers. From the experimental result can be concluded that our technique of the proposed framework achieved 25.92% with 1-best accuracy.