A Hybrid Question Answering System based on Ontology and Topic Modeling

Kwong Seng Fong, Chih How Bong
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300, Kota Samarahan, Sarawak, Malaysia.
ksfong@siswa.unimas.my

***Abstract*—A Question Answering (QA) system is an application which could provide accurate answer in response to the natural language questions. However, some QA systems have their weaknesses, especially for the QA system built based on Knowledge-based approach. It requires to pre-define various triple patterns in order to solve different question types. The ultimate goal of this paper is to propose an automated QA system using a hybrid approach, a combination of the knowledge-based and text-based approaches. Our approach only requires two SPARQLs to retrieve the candidate answers from the ontology without defining any question pattern, and then uses the Topic Model to find the most related candidate answers as the answers. We also investigate and evaluate different language models (unigram and bigram). Our results have shown that this proposed QA system is able to perform beyond the random baseline and solve up to 44 out of 80 questions with Mean Reciprocal Rank (MRR) of 38.73% using bigram LDA.**

***Index Terms*—Knowledge-Based Approach; Language Model; QA System; Text-Based Approach**

I. INTRODUCTION

A Question Answering (QA) system is an application that can provide accurate answers to the user's natural language questions. In recent years, the demand for automated QA system becomes very high. It is because it is a suitable learning platform for active and unsupervised learning. The students can seek help from QA systems when they have questions. It will be helpful if there are automated QA systems, which assist the students to learn a subject effectively and efficiently. For example, "*Ask.com*", "*Yahoo! Answers*" and "*START Natural Language QA System*", they all focus on multiple English topics. The users can input the question to the system in order to obtain the answers and information.

There are works using various approaches to improve the performance of the QA systems. For knowledge-based approach, it uses various predefined templates to form triple patterns in order to solve different question types. Sometimes, it tries to form complex triple patterns, so that it can select the appropriate answers. For text-based approach, it uses different Information Retrieval (IR) to find the answers.

In this paper, we proposed a hybrid QA system, which is a combination of both knowledge and text-based approaches. It can solve five types of the questions: factoid types, description types, definition types, reason types and relation types. Overall, our intention is to remove the complication of defining patterns and to improve the answer retrieval.

In the following section, we will present two approaches to QA systems. Section 3 will explain Latent Dirichlet Allocation (LDA). Section 4 will discuss the Q&A system architecture. Section 5 will describe a Physics ontology to support QA. Section 6 will detail out our QA system architecture. Section 7 will present the results of the QA system. Finally, section 8 will present the paper conclusion and future works.

II. APPROACHES OF QA SYSTEMS

In general, there are two types of QA systems: knowledge-based and text-based approaches.

A. Knowledge-based Approach

The knowledge-based approach has a structured knowledge base (KB). The approach embeds the keywords of the question into predefined templates to form triple patterns, which is a semantic representation of the questions to be used to extract the answers from the KB [1].

However, this approach has two factors that may affect the system performance: forming the triple pattern and retrieving the answers from the KB [1]. This is because if the system cannot form the triple pattern correctly, it cannot retrieve the correct answers from the KB. The formation of the triple pattern can be very complicated as different question types require different templates. For example, the Boolean question require the template of "*ASK WHERE ?x ?p ?y*" and the simple question require the template of "*SELECT DISTINCT ?x WHERE ?x ?p ?y*" where *?x, ?p* and *?y* are proxy variables [2].

B. Text-based Approach

The text-based approach is also known as the IR approach. This approach retrieves information from a text collection, where is unstructured. The overall process of the approach is converting the question into a query, which is a list of the keywords, and then input the query into the IR or search engine to find the relevant documents (also answers) [3][4]. With a list of relevant documents, the system ranks the most relevant documents as the answers to the questions.

However, this approach also has two factors which may affect the system performance: the formation of the query from the question and ranking of the relevant documents. It is because if the question cannot be converted into meaningful query, the system cannot find the most relevant answers, hence yielding low precision and recall [5].

III. LATENT DIRICHLET ALLOCATION (LDA)

LDA is a generative probabilistic model for a set of discrete data likes text corpora [6]. It presents the documents as a