A semantic-based Question Answering System for Indonesian Translation of Quran

Syopiansyah Jaya Putra[1], Ria Hari Gusmita[2], Khodijah Hulliyah[2], Husni Teja Sukmana[2]
[1]Information Systems [2]Informatics Engineering
UIN Syarif Hidayatullah Jakarta, Indonesia
{syopian, ria.gusmita, khodijah.hulliyah, husniteja}@uinjkt.ac.id

ABSTRACT
This paper presents a work in developing a semantic-based question answering system (QAS) for Indonesian Translation of Quran (ITQ). This research is motivated by the lacks of previous built QAS that caused by a keyword-based retrieval. Instead of keeping the retrieval method, we shifted to a semantic approach where the retrieval process is done by using a semantic similarity measurement. In doing so, we built an ontology of ITQ to get the concepts as well as verses where they appear in. We applied three factoid question types on the QAS that including Who, Where, and When. Furthermore, a weighted vector for each concept that belongs to respective expected answering type (also called as named entity group) i.e. Person, Location, and Time is generated in order to feed semantic interpreter on user question. From 222 concepts defined from the ontology, we clustered them into 77, 24, and 6 concepts for Person, Location, and Time respectively. Since we found there are some characteristics of texts in ITQ, we developed our own modules to deal with including generate the inverted index and named entity recognition. Answer extraction is conducted by applying some features extraction in order to score the answer candidates. Evaluation of the system is designed by providing two data set of question and answer where the first one is purposed to measure the effectiveness of semantic approach comparing with keyword-based retrieval and the last one aims to know system performance in regard the appearance of concepts in ITQ.

Categorized and Subject Descriptors
• Information systems → Information retrieval → Retrieval tasks and goals → Question answering.

Keywords
Semantic approach; Question Answering System; Indonesian Translation of Quran.

1. INTRODUCTION
Information Retrieval (IR) techniques have proven quite successful at locating data within large collections of documents relevant to a user's query. One of IR applications is the search engine. Google is the most popular search engine that will deliver a list of relevant documents (in form of a passage) in respect of the user query as result of the searching process. In this case, the user needs to check those relevant documents manually in order to determine which documents contain the required information. This, of course, will take more time and effort for the user. Often, the user wants not whole documents but brief answers to specific questions. For instance "*Who is the fourth Indonesian President?*", "*When do Moslem people start fasting?*", "*Where do Moslem people perform Hajj?*" Recently, a number of research projects have investigated the computational techniques needed for effective performance at this level of granularity, focusing just on questions that can be answered in a few words taken as a passage directly from a single text. These computational techniques are applied in the Question Answering (QA) research field. A QA system will allow the user to enter his/her own question in natural language and will result in the appropriate answer for the question.

In 2013, [1] conducted research into developing a Question Answering System by applying a Rule-based method on relevant documents of ITQ. This research was motivated by a lack in the previous question answering system developed by [2] which implemented an IR-based QA system. The IR-based QA system employed a search engine to retrieve relevant documents and use them to extract the appropriate answer. The weakness in [2] was caused by the search engine for Indonesian text documents that was not always able to retrieve the relevant documents and so influence the answer extraction process when producing an answer. On the other hand, research has been conducted to develop a Rule-based Question Answering system which yields satisfactory performance in producing appropriate answers [3]. The research did not utilize a search engine to locate relevant documents but employed lexical rules on all the text documents to find an answer.

The authors in [1] combined standard QA system architecture (using a search engine) and the rule-based method in order to increase the accuracy of the QA system. Based on the results in [1], it was surprising that the enhanced system was not able to achieve good performance and was even worse than the original system [1] that only used one technique to produce an appropriate answer. Analysis of the result showed that failure came from the inability of the search engine in processing Indonesian text documents and non-compliance with lexical rules. Furthermore, in an IR-based QA system, the keyword search lacks a clear specification of the relations among entities and so gives a contribution in delivering irrelevant documents.

Recently, another kind of QA system for Arabic Quran has been developed that strengthens its architecture by the addition of a Knowledge Base (KB). A knowledge-based question answering (KB-QA) computes answers to Natural Language (NL) questions based on existing knowledge bases (KBs) [4]. Most previous systems tackle this task in a cascaded manner whereby in the first

iiWAS '16, November 28-30, 2016, Singapore, Singapore

DOI: http://dx.doi.org/10.1145/3011141.3011219

step the input question is transformed into its Meaning Representation (MR) by an independent semantic parser. Then, the answers are retrieved from existing KBs using generated MRs as queries [4]. In the previous work, we used TF-IDF for generating a weighted vector to obtain the optimal concepts in ITQ ontology. By using a weighted vector, a KB-QA will retrieve a semantically related document to be used in answer extraction.

In aiming to improve the performance of previous QAS for ITQ in [1,2], we adopted a semantic approach to the QA system. This paper will describe a work on developing QAS for ITQ using semantic approach.

2. PREVIOUS WORKS

Wardani [5] presented there are two types weighted schemes: unsupervised term weighting schemes (UTWS) and supervised term weighting schemes (STWS). TF–IDF, short for term frequency–inverse document frequency, is one of some methods that it is often used as a weighted factor in UTWS. Moreover, TF-IDF has been used in order to provide the main concepts of the domain for the ontology construction.

In that respect, [6] presented research that there are three levels approach for Arabic question answering system. This work is composed of a keyword-based level relying on Query Expansion (QE) using Arabic WordNet (AWN) semantic relations, a structure-based level integrating the Distance Density N-gram (DDN) model and a semantic-based level considering the representation of meaning in both the question and the corresponding passages using the Conceptual Graphs (CGs) formalism and the comparison based on the semantic similarity score

The other previous research in Question Answering for the Indonesian language and ITQ has been widely applied [7] Proposed a machine learning approach for Indonesian Question Answering system. These systems apply the SVM as a machine learning algorithm. The question with *"kapan"* (when) as the interrogative word is always a date question. [3] Developed a Question Answering using a rule-based method on the text of the Quran in The Indonesian language.

3. SYSTEM ARCHITECTURE

The following figure depicts how the semantic-based question answering system was developed based on the following process as shown below;

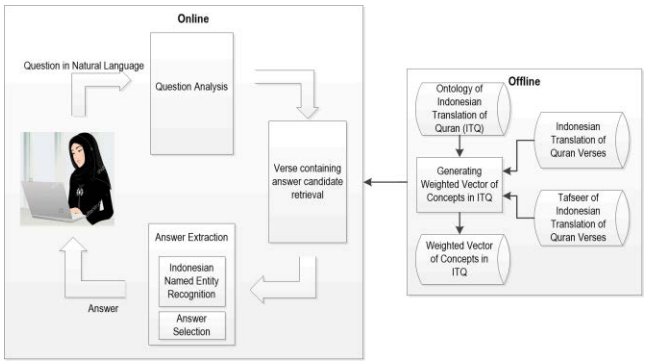

Figure 1: Architecture of Semantic-based QAS for ITQ

As seen in Figure 1, there are two phases to conduct a process to generate Semantic-based QAS for ITQ that we called as Offline Phase and Online Phase.

Offline Phase is purposed to provide all data resources needed for processing at Online Phase. Those data resources are including Ontology of ITQ, ITQ, and weighted vector for concepts in ITQ. Each data resource is gathered in particular way especially for weighted vector for concepts in ITQ that there was a processing task to generate it by employing two others data. This all task in the online phase is explained in [10]. Resulted weighted vector for concepts in ITQ is utilized to do verse retrieval in the Online Phase.

Online Phase has three components which are Question Analysis, Verse Containing Answer Candidate Retrieval (VCACR), and Answer Extraction (AE). The question in natural language form taken from the user is then analyzed in Question Analysis component so that it produces an expected answer type. This output is applied by VCACR in order to retrieve verses that contain answer candidate base on semantic similarity measurement. Finally, all retrieved verses are scored in AE component by implementing several rules.

3.1. Question Analysis

At the beginning, the system will take user question in order to transform the question into a weighted vector and output the expected answer type. The weighted vector will be used in VCACR stage in fulfilling semantic similarity measurement between question and each verse to retrieve relevance verses. This transformation process is called as semantic interpretation[8]. We implement the same way as in [9] to get a weighted vector of the question. These following steps are to conduct semantic interpretation on the question that we derived from [9].

1. Given a text fragment of the question, ranks all the ITQ concepts by their relevance to the fragment. We set each of meaningful term in the question as a fragment. We applied fragment in single and phrase form.
2. Given a text fragment, we first represent it as a vector using TFIDF scheme.
3. For each text fragment, retrieves corresponding entries from the inverted index, and merges them into a weighted vector of concepts that represents the given text.
4. Let P = $\{w_i\}$ be input text, and let v_i be its TFIDF vector, where v_i is the weight of word w_i.
5. Let id_j be an inverted index entry for word w_i, where id_j quantifies the strength of association of word w_i with ITQ concept c_j, $\{c_j \in c_1, \ldots, c_N\}$ (where N is the total number of ITQ concepts).
6. Then, the semantic interpretation vector V for text T is a vector of length N, in which the weight of each concept c_j is defined as $\sum_{w_i \in T} v_i \cdot k_j$.
7. Entries of this vector reflect the relevance of the corresponding concepts to text T.

Expected answer type of the question is determined base on question types applied in the system. Since this research is purposed to increase the performance of our previous QAS, we used the same question types that are Who, Where, and When. This following table shows the expected answer type for each of those question types.

Table 1 Question type and expected answer type list

Question type	Expected Answer Type
Who	Person
Where	Location
When	Time

3.2. Verse Containing Answer Candidate Retrieval

This stage is giving a big different between the standard architecture of QAS as used in [1,2] and semantic approach in the QAS. Instead of conducting a keyword-based search in the retrieval process, we applied semantic similarity measurement to get relevance verses. In doing so, we implemented cosine similarity method to measure the similarity between question and each verse in ITQ. There are several preliminary tasks needed to accomplish this process as listed below;

1. Have each concept and verses where it appears in associated by defining a weighted vector;
2. Generate an inverted index for each term in ITQ;
3. Measure the semantic similarity between question and each verse in ITQ.

Weighted vector generation of each concept in ITQ has been done by using TFIDF method as described in [10]. In regard to generating the inverted index for each term in ITQ, we developed our own module to proceed it. This is because in our previous researches, Lucene as a tool to do searching as well as indexing process did not perform well as it remains an over Stemming. Building this module brought us to some discovery things in ITQ as explained as follows:

1. There is inconsistency on term writing. This is can be seen on term ‘Ad which at some verses it appeared as ‘Aad. To deal with this condition, we chose the form that appeared more and also written in other translation of Quran. For term ‘Ad, we used ‘Aad and applied it in the whole ITQ content.
2. There is a typo mistake on term writing. It is shown mostly on the term that using repetition such as orang-orang, benar-benar, etc.
3. Previously, we also used Tafseer of ITQ and combined it with the related verse. After the preprocessing task was done, we found that there were many noises came along and caused some problems. Those noises are transliteration marks, special characters that represent a comment, and even some terms exist in a different literal than its appearance in ITQ. This is furthermore led us to withdraw Tafseer of ITQ from our Corpus.

This following is sample of our inverted index for each unique term in ITQ:

```
al ahqaaf,46_21.txt
al lata,53_19.txt
al masih isa
al masjidil aqsha,17_1.txt
al quran,11_17.txt,41_44.txt,42_52.txt,2_91.txt,41_4
al uzza,53_19.txt
alah,64_9.txt,4_92.txt,3_165.txt,12_50.txt,40_9.txt,
alaikum,13_24.txt,16_32.txt,7_46.txt,6_54.txt
alam,23_80.txt,3_190.txt,25_59.txt,28_73.txt,73_20.t
alami,36_68.txt,27_88.txt,4_3.txt,23_18.txt,2_136.tx
alangkah,19_38.txt,6_31.txt,56_9.txt,56_8.txt,54_16.
```

Figure 2 Sample of inverted index for unique term in ITQ

In Figure 2, each term is written at the first following by a list of document name where the term appears in. Document names are separated by using comma character.

Semantic similarity measurement is applied by using cosine similarity method. It is conducted by following this formula:

Cosine Similarity(Question,Verse) = Dot product(Question, Verse) / ||Question|| * ||Verse||

3.3. Answer Extraction

After having semantic relevance verses, this component will apply named entity recognition and feature extraction to select the best verse and extract the answer. Named entity recognition is done by putting a proper tag at term base on its named entity class. This named entity class is related to expected answer type. As mentioned before, we applied three question types and that the expected answer types are Person, Location, and Time, we also utilized them as named entity classes. We developed our own named entity recognition module to fit with all characteristics in ITQ. This module will be used to annotate each verse in ITQ as well as user question. In accommodating named entity recognition, we determined the member of each named entity group first by analyzing all verify leaf concepts in the ontology of ITQ manually. From our ontology, we got 222 concepts. We selected 77, 24, and 6 concepts that belong to Person, Location, and Time named entity respectively. By referring to this named entity members, named entity recognition is done both on question and verses where the annotation is set like XML tag format in that each tag is completed with a closing tag. This is a sample of named entity recognition on a verse.

```
dan ingatlah <Person>hud</Person>
saudara kaum <Person>aad</Person> yaitu
ketika dia memberi peringatan kepada
kaumnya di <Location>al
ahqaaf</Location> dan sesungguhnya
telah terdahulu beberapa orang pemberi
peringatan sebelumnya dan sesudahnya
dengan mengatakan janganlah kamu
menyembah selain <Person>allah</Person>
sesungguhnya aku khawatir kamu akan
ditimpa azab hari yang besar
```

Figure 3 Sample of Named Entity Recognition on a Verse

Feature extraction is the last step to be implemented before extract the answer. It will be applied on semantic relevance verses in order to get the probability of correctness of a question and particular verse. This kind of verses is called as answer candidate. Afterward, the verse that has the highest probability of correctness is returned as the answer. We determined a set of features to be used in calculating probability of correctness that derived from [10] as follows:

1. Maximum number of matched words between the input question and the answer candidate;
2. The type of the question’s expected answer if it matches with the extracted named entity in the answer candidate;
3. The maximum count of named entity types that occurred in the question occurring in the candidate answer.

4. The minimum distance between matched terms in the passage.

4. Evaluation

Evaluation of our QAS is done by giving predefined questions that represent each expected answer type. Each of questions is completed with the answer as well as verse location where the answer found in. This dataset question and answer is determined manually and to be a benchmark for the system. We provided two groups of dataset that including dataset used in [1] and the one that defined based on concepts in ITQ. The first dataset is purposed for doing a comparison between system model adopted in [1] and semantic approach in this research. The last dataset aims to know system performance in regard to the existence of concepts in ITQ.

5. CONCLUSION

We have presented a work in developing a semantic-based QAS for ITQ. Rather than using a keyword-based retrieval, semantic approach brings the system to retrieve semantic relevance verses based on a measurement process by using Cosine similarity method. Text processing on ITQ was not only resulted data resources but delivered several facts on ITQ that obviously become a valuable information to the future research in this field. Currently, we are going to do evaluation process by following evaluation scenario that has been determined.

6. ACKNOWLEDGEMENT

This work was supported by the Directorate of Islamic Higher Education, Ministry of Religious Affair of the Republic of Indonesia for International Collaborative Research Cluster Code No KNI/59/2015. We would like also thank for this collaboration to Ismail Khalil, the Institute of Telecooperation, Johannes Kepler University Linz, Austria.

7. REFERENCES

[1] Gusmita, Ria Hari, et al. "A rule-based question answering system on relevant documents of Indonesian Quran Translation." Cyber and IT Service Management (CITSM), 2014 International Conference on. IEEE, 2014

[2] Zidny, Nafán, and Gusmita, Ria Hari,, "Developing an Indonesian Question Answering System about Khulafaur Rasyidin History", in Proceedings of The 1st International Conference on Cyber & IT Service Management in Conjunction with the ITIL v.3 Workshop, Training and Certification, Bandung, Indonesia, 2012.

[3] Anggraeny, Meynar Dwi. "Implementasi Question Answering System Dengan Metode Rule-Based Pada Terjemahan Al Qur'an Surat Al Baqarah." (2007).

[4] Bao, Junwei, et al. "Knowledge-based question answering as machine translation." Cell 2.6 (2014).

[5] Wardani, Dewi Wisnu, and Wen Hsiang Lu. Finding Structured and Unstructured Features to Improve the Search Result of Complex Question. Diss. National Cheng Kung University, 2009.

[6] Cheddadi, Abdelkhaleq. Three-levels Approach for Arabic Question Answering Systems. Diss. Ecole Mohammadia d'Ingénieurs, 2014.

[7] Purwarianti, Ayu, Masatoshi Tsuchiya, and Seiichi Nakagawa. "A machine learning approach for indonesian question answering system." Artificial Intelligence and Applications. 2007.

[8] Abdelnasser, Heba, et al. "Al-Bayan: an arabic question answering system for the holy quran." ANLP 2014 (2014): 57.

[9] Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing semantic relatedness using Wikipedia-based explicit semantic analysis." IJcAI. Vol. 7. 2007.

[10] Putra, Syopiansyah Jaya, et.al, "A semantic-based Question Answering System for Indonesian Translation of Quran", final report on International Collaboration Research Cluster Code No KNI/59/2015, funded by Ministry of Religious Affairs Republic of Indonesia, 2015.

[11] Alrehaili, Sameer M., and Eric Atwell. "Computational ontologies for semantic tagging of the Quran: A survey of past approaches." LREC 2014 Proceedings (2014).

[12] Ta'a, Azman, et al. "Al-Quran themes classification using ontology." (2012): 383-389.