

Eliminating Unanswered Questions from Question Answering System for Khulafaa Al-Rashidin History

Muhammad Zidny Naf'an¹, Dewi Erla Mahmudah², Syopiansyah Jaya Putra³, Asep Fajar Firmansyah³

Department of Informatics¹, Department of Informatics Engineering², Department of Information Systems³

ST3 Telkom¹, STMIK Widya Utama², Islamic State University of Syarif Hidayatullah³

Purwokerto-Indonesia^{1,2}, South Tangerang-Indonesia³

zidny@st3telkom.ac.id, mdewierla@gmail.com, syopian@uinjkt.ac.id, asep.airlangga@uinjkt.ac.id

Abstract—This paper discusses some attempts conducted to improve performance of Question Answering System for Khulafaa Al-Rashidin (Caliphs) history (called as QAKH). Experiments done on QAKH in 2012 showed that only 61,67% questions got a correct answer. The one contributed in this achievement was lack of Indonesian stemming process that implemented by utilizing Lucene library. It was found that Lucene did over stemming on some Indonesian word in indexing phase, so that it was delivered impacts on passage retrieval and answer extraction as well. We tried to implement two approaches in order to solve that problem. As the first, we used other library for doing stemming process for Indonesian words that developed by Information Retrieval laboratorium in University of Indonesia. In any case if there is no answer delivered, we shifted to the second approach where there is no stemming process applied. Result of evaluation demonstrated a better achievement where we got 66,67% correct answers. Analysis of each fold of experiments also discussed.

Keywords—*eliminating unanswered question; question answering system; khulafaa al-rashidin history;*

I. INTRODUCTION

History is something important for human life. The people can learn about life from the predecessors. First President of Indonesia, Ir. H. Soekarno had made the remark known as “JAS MERAH”, short form of “Jangan sekali-kali melupakan sejarah” (Never forget history). Muslims need to know the history of the predecessor people. Moreover, Muslims need to know the group of people which was studying with the Prophet Muhammad PBUH (*Peace be Upon Him*) directly, known as the Sahaba (companions of the Prophet). The Sahaba is a group of the best Muslims, The Prophet PBUH said: "The best of my Ummah (group of Muslims) are the people whose contemporaries with me" (Hadith Narrated by Al-Bukhary).

A question answering system has been developed for gaining information about Khulafaa al-Rashidin history in 2012 (called as QAKH) [1]. Corpus used in the system contained information about the Prophet's Sahaba who were asked to be the caliph after the Prophet. The Four Sahaba who became chalips were Abu Bakr Al-Sadiq, Umar ibn al-Khattab, Uthman ibn 'Affan, and Ali ibn Abdul Muttalib - may Allah blessing them. Unfortunately, QAKH did not get

it's best performance where from 60 submitted questions, it only succeed delivering 61,67% correct answer and 21,67% questions left with no answer. The one contributed in this achievement was lack of Indonesian stemming process that implemented by utilizing Lucene library. It was found that Lucene did over stemming on some Indonesian word in indexing phase, so that it was delivered a bad impacts on passage retrieval and answer extraction as well.

This paper describes some efforts aiming to increase performance of QAKH in term of eliminating number of unanswered questions. We employed two approaches to boost number of answered question. The first approach was utilized other Indonesian stemming process tool developed by Information Retrieval Laboratory in University of Indonesia. For any case if this method failed in providing an answer, we move to the second line where there is no stemming process applied.

In section II we briefly discuss previous works in question answering system development conducted for Khulafaa al-Rashidin history, before outlining proposed approaches to eliminate the unanswered questions in section III. Subsequently, we present experiment results and analysis in section IV. A summary and further tasks are described in section V.

II. PREVIOUS WORK

Question Answering System (QAS) on the Islamic field has been explored in several researches. Naf'an, et al [1] built a QAS for Khulafaa al-Rashidin history. It employed standard architecture of QAS where consists of Question Analyzer, Passage Retrieval, Named Entity Recognizer, Passages Scoring, and Answer Extraction. Since they did not find any digital document of Khulafaa al-Rashidin history, corpus was produced by typing reference book's content manually. Each resulted document represented one chapter on the book. From each document, some passages were generated in order to have retrieval process run effectively. Passage retrieval was conducted by applying Lucene library and algorithm adopted from [5] was employed for extract the answer. On the evaluation, it was shown that the system only able to returned 61,67% correct

answers. The rest supplied answers were ranging from Unsupported, Inexact, and Wrong answer.

Aiming to improve performance of QAS built in [1], some efforts were carried in 2014 [8]. Those were adding new components on the architecture and applying new method for solving previous weaknesses. There were two new components called as Question Grammatical Checker and Usage Knowledge which functioned at accommodating incorrect question and keeping pair of processed question and answer in order to have faster processing time respectively. Furthermore, over stemming granted from Lucene was handled by utilizing an Indonesian Big Dictionary on stemming process. Surprisingly, even equipped with more resources the system did not get its expected results and had a lower achievement than [1]. It only succeed to had 32% questions with correct answers.

Gusmita, et al. employed a new architecture to construct a QAS for Indonesian translation of the Qur'an. New architecture designed from a combination of standard architecture of QAS and Rule-based method of QAS proposed by [7]. Since there is a drawback on passage retrieval at standar architecture[1,8] and conversely [9, 10] had successfully applied a rule-based method on QAS where they got number of correct answer more than 85%, combination of both architectures was chosen to attain good performance. Unfortunately, this one-step-forward approach still failed to enhance system performance as it resulted only 53.33% correct answered question.

III. PROPOSED FRAMEWORK

Analysis of experiment results on QAKH in [1] shown several things caused some questions left without answers as follows:

1. Search engine could not find passages that was relevant to the query.
2. System could not find query's keyword on the passage that was returned by search engine.
3. System could not find the word that has the same named entity with type of expected answer on the returned passage

Those analysis result bring us to an initial hypothesis where they were an impact of over stemming contributed by Lucene on stemming process. Over stemming give a probability to have an incorrect index for a passage. This incorrect index will let the search engine fails to find relevant passages or retrieved irrelevant passages and so that QAS could not find query's keyword or word with expected answer type named entity. To deal with this condition, our proposed framework focuses on stemming process. We designed two scenarios to meet our expectation in eliminating unanswered questions as listed below:

1. As the first one, we substituted Lucene's task in doing stemming process by utilizing an Indonesian stemming tool (<http://fws.cs.ui.id/Stemmer/>). Result of stemming process will be employed by Lucene to do indexing and passage retrieval as well.

2. For any question failed to be answered in the first scenario, we do reprocess on that question where passage retrieval is implemented without stemming process.

To support the experiments, we used the entirely component on the QAS architecture applied in [1] including question analyzer, passage retrieval, passage scoring, and answer extraction as depicted on Figure 1 as follows:

1

Figure 1. QAS Architecture in [1]

Applying our framework, in the preprocessing phase we generated two kinds of passage's index. The first index produced through stemming process that applied by using Indonesian stemming tool. Furthermore, the second index resulted without stemming process. We used Lucene in version 6.0.1 for indexing and passage retrieval purpose. In order to have faster running time on answer extraction phase particularly on passage scoring, we created a new passage collection where they were a fruit of stemming process. This following is sample of stemmed passage:

nasab temu nasab Rasulullah saw. kakek nama Salma binti Shakhr bin Amir bin Ka'ab gelar Ummu Khair . Imam An-Nawawi kitab tahdzib al-asma' kata "nama Abu bakar ash-siddiq masyhur Abdullah

On the question analyzer, we extracted several data to be utilized on the next component. They were keyword entity, keyword, and expected answer type. We also produced Boolean query where AND operator took a place. This Boolean query was used on passage retrieval component where based on analysis conducted in [1], there were 20 top passages returned. Lucene 6.0.1. was also benefited in this retrieval. Table I shows sample of result from question analyzer component.

TABLE I. Sample of question, boolean query, keyword, and keyword entity

Question	:	Siapakah nama lengkap abu bakar?
Boolean Query	:	nama AND lengkap AND abu AND bakar
Keyword Entity	:	[abu bakar]
Keyword	:	[nama, lengkap]

Once top 20 passages were delivered, they would be scored by adapting scoring rules in [5]. Ranking process is applied after we got each passage scored where the highest passage score was on the top position. On the answer extraction, we implemented several tasks to get the answer as outlines below:

1. System will check each passage whether it contains any or all keywords and contain the word entity appropriate with the expected answer types or not.
2. If it contains both of them, then calculate the distance from keyword to word entity in the passage.

3. The next list of the answers candidate will be sorted by distance, passage's score from Passage Scoring, and passage's score from search engine.
4. Answers candidate at the top list is the answer to the user question. Furthermore, this answer will be displayed to the user.

IV. RESULT AND DISCUSSION

The answer categories referring to [1] are correct, incorrect, unexact, unsupported, and unanswer. Unsupported category mean the system returns the correct answer, but obtained from documents that are not relevant with the question. Unexact category mean the system returns the answer and the answer string which is more than the actual answer, or lose some of the answer and extracted from the documents relevant to the query [1]. The result of our research is as follows:

TABLE II. Result from our Framework

Categories	Result	Percentage
Correct	40	66,67%
Unsupprted	1	1,67%
Unexact	2	3,33%
Incorrect	17	28,3%
Unanswer	0	0,00%

From Table II, it can be seen that there were an increase on number of correct answer and surprisingly we were success to have zero number on unanswer question. This result shows a good achievement of our proposed framework as it is totally different with result of [1] as follows:

TABLE III. Result of Previous QAS [1]

Categories	Result	Percentage
Correct	37	61,67%
Unsupprted	1	1,67%
Unexact	1	1,67%
Incorrect	8	13,33%
Unanswer	13	21,67%

We provides a line chart to demonstrate result comparison between our framework and previous system on Figure 3.

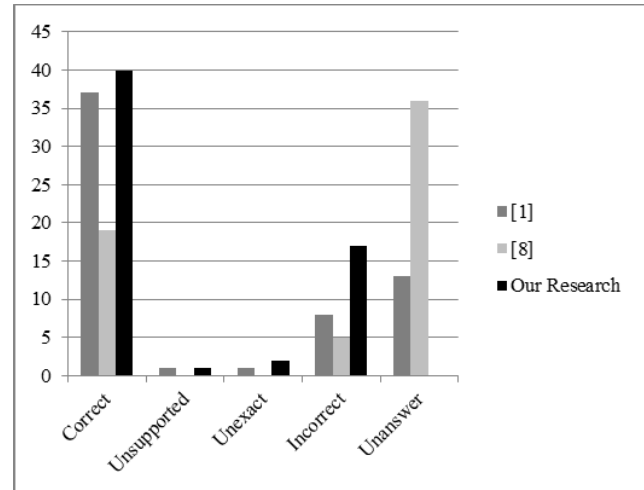


Figure 3. Comparison of our Framework with Previous System

From 13 unanswered questions that finally able to be answered, we grouped them base on the answer correctness. We got 4 answers were correct, 1 unexact answer, and 8 incorrect answers as listed on Table IV.

TABLE IV. Status Unanswered Questions in [1] in our Framework

Categories	Result	Percentage
Correct	4	30,77%
Unexact	1	7,69%
Incorrect	8	61,54%

The example of unanswered question on [1] and correct answered in this research is the fifth question: “Siapakah nama ibu Umar?”. Based on [1] the system could not find keyword “nama ibu” on passage that returned by search engine. With rule used on the Answer Extraction from 2nd Framework, the system can find answers candidates and managed to get correct answer, it is: “Hantamah binti Hisyam bin Mughirah”.

The example of unanswered question on [1] and unexact answered in this research is the 11th question: “Siapakah nama ibu ali?”. The cause of research [1] did not find the answer to this question same with the fifth question above. System in this research could not restore answer precisely because the complete answer truncated when performing pre-processing (stemming content passage), so that the system only restore the answer “Fathimah”, whereas the correct answer is “Fatimah binti Asad bin Hasyim”.

While the example of unanswered question on [1] and incorrect answered in this research is the 8th question: “Siapakah putri Rasulullah yang menikah dengan Utsman?”. The system could not find the correct answer because an error occurred on the Answer Extraction, that is the absence of rule to check whether all keyword found in passage or not. Answer Extraction will measure distance any or all keyword with entity word. On a passage that generates answers to this question, that is passage Ali Bin Abu Thalib-Tokoh-Tokoh yang Meninggal di Zaman Ali--3--3, only contain keyword “menikah” and did not contain keyword

“utsman”. The answer’s system with answer candidates has shortest keyword distance than others.

V. CONCLUSION AND RECOMMENDATION

Based on the results obtained in the experiment, question-answer system was successfully answer all the questions. However, not all the questions were answered correctly. when compared to the study [1], our proposed framework has better results, with the percentage of questions correctly answering system is 66.67%. While in [1] the percentage of correct answers was 61.67%. Further research will be focused on how to handle incorrect answers.

VI. REFERENCES

- [1] M. Z. Nafan, R. H. Gusmita, and M. T. Rosyadi, "Developing an Indonesian Question Answering System About Khulafaur Rashidin History," in *1st International Conference on Cyber and IT Service Management*, 2012, pp. A-41.
- [2] Al-Suyuthi, *Tarikh al-Khulafa: Ensiklopedia Pemimpin Umat Islam dari Abu Bakar hingga Mutawakil*, Fahri, Ed. Bandung, Indonesia: Hikmah, 2010.
- [3] B. Yatim, *Sejarah Peradaban Islam*. Bandung, Indonesia: Rajawali Pers, 2010.
- [4] Laboratorium Perolehan Informasi Fak. Ilmu Komputer, Universitas Indonesia. (2016, Jul.) Question Answering. [Online]. HYPERLINK "<http://ir.cs.ui.ac.id/qa.php>"
<http://ir.cs.ui.ac.id/qa.php>
- [5] S. H. Wijono, I. Budi, L. Fitria, and M. Adriani, "Finding Answers to Indonesian Questions from English Documents," Universitas Indonesia, 2006.
- [6] R. Ellen and T. Michael, "A Rule-Based Question Answering System for Reading Comprehension Tests," in *Workshop on Reading Comprehension Tests as Evaluation for*, 2000.
- [7] A. H. Lestari, "Improvement of Khulafaur Rasyidin’s Question Answering System," Islamic State University of Syarif Hidayatullah Jakarta, Dept. of Science and Technology Bachelor Thesis 2880 TI s, 2014.
- [8] R. H. Gusmita, et al., "A Rule-Based Question Answering System On Relevant Documents Of Indonesian Quran Translation," in *International Conference on Cyber and IT Service Management*, South Tangerang, 2014, pp. 104-107.
- [9] M. D. Anggraeny, "Implementasi Question Answering System Dengan Metode Rule-Based Pada Terjemahan Al Qur’an Surat Al Baqarah," IPB Undergraduate Thesis, 2007.
- [10] L. C. Rosiana, "Question Answering Sistem pada Terjemah Juz Amma Menggunakan Metode Rule Based," Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim Undergraduate Thesis, 2012.