Construction of Question Answering System for Factoid and Non-Factoid Type Question for History of Khulafaur Rasyidin

1st Nurhayati[1], 2nd H. A. Nadisah[1], 3rd A. H. Setianingrum[1],4th Busman[2]
{nurhayati@uinjkt.ac.id[1], nadisa@mhs.uinjkt.ac.id, anif.hanifa@uinjkt.ac.id[1],
busman.sjam@gmail.com[2]}

UIN Syarif Hidayatullah, Department of Informatics, Jakarta, Indonesia[1], UIN Syarif Hidayatullah, Department of Management, Jakarta, Indonesia[2]

Abstract. Question answering system is one of sub-fields information retrieval that integrates information from multiple sources as document collection whether local or web to provide concise answer to specific question which is entered by the user. This question answering system is about history of Khulafaur Rasyidin which is expanded from another researcher with adding types of question word; such as: "what, why, and how". So, this question answering system provides question words: "who, when, and where" as factoid question and "what, why, and how" as non-factoid question. This question answering system consists of several processes: question analyzer, document retrieval, and answer extraction. The question analyzer aims to classify question to obtain answer type, clue word, and query. This query will be used in document retrieval to obtain document candidate which contain exact answer by using search engine which is called Lucene. The sources that are used came from previous research. It has been we reviewed and corrected. Answer extraction aims to obtain exact answer which is desired by the user. Answer extraction for factoid question is performed by calculating distance between answer candidate to all queries from the question in sentence answer and for non-factoid question is performed by sentence pattern which has priority value for each sentence. System testing is performed by giving 30 question words for factoid question and non-factoid question. The evaluation of answer result refers to another researcher which has 5 categories, namely:" true, inappropriate, not supportive, unanswered, and wrong". The result of the system evaluation shown factoid question has high performance to obtain exact answer and non-factoid question has low performance.

Keywords: Information Retrieval, Question Answering System, Factoid and Non-Factoid Type Question, The history of four caliphs (Khulafaur Rasyidin)

1 Introduction

Nowadays, information technology has developed rapidly. Advancement of technology has brought human race in the new cultural pattern. Nearly, every human activity uses media information technology to solve a problem. It is caused by a number of inventions and innovations of the information technology that doesn't stop growing. Based on this condition,

ICONQUHAS 2018, October 02-04, Bandung, Indonesia

DOI 10.4108/eai.2-10-2018.2295538

the impact makes information communication technology has substantial position on daily activities.

Construction of information technology gives a major influence on human. It will look wasting if it's not used to look for information, particularly in the Islamic world. Therefore, it takes an integrated system that can be created between technologies and Islamic, such as by utilizing question-answering system that related to Islam. It research based on "Question Answering System in Indonesia Language for Khulafaur Rasyidin History" [1].

The Question types which are handled by question answering system are divided into 5 types, there are: factoid, non-factoid, yes-no, list, and opinion [2]. Question answering system that has been developed by [1] only applied three types of question word, namely: "who", "when", and "where". According to [2] those three types of question word are also called factoid question, where factoid question is a question which has answer that contains phrase such as: "people, location, organization, date, number", and other short answer types. Meanwhile according to [3], factoid question is a question which based on fact and has short answer. So far, there isn't elaboration of question answering system which has been developed by [1] yet. Meanwhile, several other researchers are used as a method to develop this system by us. Therefore, we will perform system elaboration in order to handle non-factoid question which often is used to understanding something. Based on those matters above, the impact of system elaboration can make people who will use this application easer to look for information about Islamic history, especially Khulafaur Rasyidin than have to read the detail books.

2 Literature Review

2.1 Question Answering System

Question answering system is one of sub-fields of Information Retrieval [4]. Question answering is a technology to find information from a big text based on a given question [3]. The question answering system virtualization is that user to state his or her information need in a more specific and natural form. It is like a natural language question and that it doesn't return full documents. They have to be skimmed by the user to determine contains. The answer is short text and phrases [5].

2.2 Architecture of Question Answering System

A prototypical question answering system of general architecture displayed the main components. The prototypical system has four components: question analysis, document retrieval, document analysis, and answer extraction. Each of these components discussed below [5]:

2.3 Analyze Question

The first phase is to analyze the question. It included a morpho-syntactic analysis of the question. The question is also classified to determine what it is asking for, in example whether it is asking for a date, a location, the name of person, etc [5]. The aplication in Indonesia language.The question words are formulated by 5W+1H, that is: "w*hat, where, who, when, why,* and *how".* They shown in the table: [6]

Table 1. Types of Question Words

Question	Meaning
What	State of Thing
Where	State of Place
Who	State of Person
When	State of Time
Why	State of Causal
How	State of Condition or Manner

A retrieval query is formulated which is posed to the retrieval component. The question analysis component must formulate a request based on the retrieval component as an addition to classifying the question. Each first word is normalized to its morphological roots. Usually, this is done using rule-based stemmers, such as the Porter stemmer or by looking for morphological roots in a machine-readable dictionary. The morphologically normalized words are used to submit request the query to the retrieval engine.

2.4 Document Retrieval

The retrieval component is generally a standard document retrieval system that identifies documents. They contain terms from the request given. The retrieval component returns a set or list of document ratings. Then it analyzed by the document analysis component.

The retrieval component returns both a series of non-sequential documents that may contain answers or a list of document order depending on the actual retrieval machine used. These documents are ranked in relation to their possibility of containing answers.

Document retrieval is only one component of the overall question answering process. Its effectiveness is very important for the overall performance of the question answering system. The document retrieval fails to return any document analysis and the answer selection component will also fail to identify the correct answer [5].

2.5 Document Analysis

The processes of Documents Analysis are: The first; it takes as an input document that contains answers to the original questions, together with the question analysis component to be generated. Then do extracting a number of candidate answers which are sent to the answer selection component.

The set of entities named are: person's name, organization, date, location, temporal, spatial distance, etc. it must be related to the information needs expressed by the question, to consider it, if a phrase is the right type. Potential answers or candidate answers.

The document analysis component passes through the list of candidate answers to the answer selection component. Each candidate's answer is related to the question. Example: it is due to analyzing syntactic structures, applying pattern matching, lexical chaining, or proximity constraints. [5].

2.6 Answer Selection

The final component chooses a phrase that is most likely the answer to the original question from the phase of the candidate's answer derived from the document analysis component. The choice answer component does not have to return one final answer to the

user, but can also return a list of ranking answers, where the answers are ordered in connection with the trust that the system has in each.

Document analysis components do not provide candidate answers or only candidates. It is only related to questions with proximity and only related to low frequency. Finally the answer selection component can decide to jump back to the question analysis component and try to reformulate the retrieval question by adding the candidate's answer [5].

3 Analysis and Design The Question Answering System Architecture

Question answering system architecture that will be used for this study is as follows:

The system has three main elements, there are: input, process, and output. The architecture above explains that the processes are started from input that is entered by the user which called question, and then passed through several processes that are performed by search engine which produce the expected output information that is suitable with the user needs that is obtained from document in which is stored with (.txt) file form that related to history of Khulafaur Rasyidin.

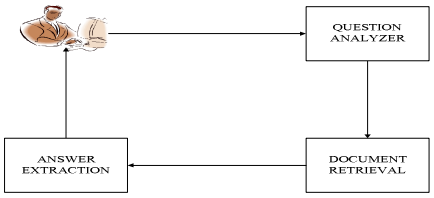

Figure 1. Architecture of Question Answering System on History of *Khulafaur Rasyidin*

The question answering system divide into three components; first component is called question analyzer. User enters a question as input in question analyzer component, while the output from this component will be an input for the second component that is called document retrieval, in which document is used in this component is about History of Khulafaur Rasyidin that is obtain from previous research by [1]. The output that is retrieved by document retrieval component is used in the third component; it's called answer extraction to obtain exact answer to question which is entered by the user.

The component of the question answering system architecture is used:

3.1 Question Analyzer

The scenario of this component is user enters a question, and then question will be processed by question analyzer to be analyzed. Question analyzer will analyze the question which entered by the user to obtain expected answer type, clue word, and query as output from question analyzer component.

Expected answer type and clue word are obtained by examining a question word which is entered by the user by doing string matching to question. Question word that is entered by the user can be classified into two types: factoid question and non-factoid question. Factoid question is question which looks for short answers and based on fact such as: "Who",

“When”, and “Where”. Whereas non-factoid question is question which looks for explanation or definition answers such as: “What”, “Why”, and “How”. The following is table of question words based on Expected Answer Type (EAT) classification rule [7] which has been we modified:

Table 2. Eat Classification Rule

EAT	Question Word	Clue Word
Factoid		
Person	*Who, Who Is*	-
Location	*Where, Where Is*	-
Time	*When, When Is*	-
Non-factoid		
Definition	*What, What Is*	*Is,, that is, meaning, , known, meaningful, is, called, , intended*
Reason	*Why*	*Because, therefore, causes, causes, consequences, causes, are caused, therefore*
Method	*How, What about*	*By the way, as, with, processes, procedures*

If question word is found, *question analyzer* component will store the appropriate *answer type* to the question word.

To explain the performance of question answering system in *question analyzer* component in order to determine *answer type* and *clue word*, the author makes two scenarios. First scenario is the example for *factoid* question. The user enters a *factoid* question to the system, such as:

"What is the full name of Abu Bakar?"

from a *factoid* question above, *question analyzer* will analyze the question until obtain a question word *“Who”*. Based on EAT classification table, question word “*Who*” refers to PERSON *expected answer type*. Meanwhile, second scenario, a question which is entered by the user is *non-factoid* question, such as:

“Why in the year of Uthman's goverment was the year of nosebleeds?”

from the question above, *question analyzer* will analyze the question to obtain a question word *“Why”*. Based on EAT classification table, question word *“Why”* refers to REASON *expected answer type* and has *clue word “"Because", "therefore", "cause", "cause of ", "effect", "causes ", "caused", or "therefore".*

karena”, “oleh karena itu”, “penyebab”, “sebab”, “akibat”, “menyebabkan”, “diakibatkan”, or *“oleh sebab itu.*

In addition to the processing of determining *expected answer type* and *clue word* in the *question analyzer* component, there is also the *process* of *query extraction*. The *query* is words that can be used to describe the content of question and its words in a question or words that is related to word in question. *Query* will look for *document* which is approximated contains answer to question which is entered by the user. In other words, wrong *query* can produce wrong relevant *sentences* so it doesn’t produce exact answers to question.

In *question analyzer* component, the process of *query extraction* is performed by removing question word, removing question mark, and removing *stopword* to the question. *Stopword* list which is used in this system is obtained from [1] which was modified by the

author. *Stopword* list which is used in this question answering system are: *who, at, and, or, with, etc.*

Based on the example of *factoid* question which is entered by the user, so *query* which is obtained from the question is:

"Full name of Abu Bakar"

whereas, for *non-factoid* question, *query* which is obtained from the question is:

"The year of Uthman's goverment was the year of nosebleeds".

3.2 Document Retrieval

Document which is used in this research based on previous research by [1], It has been reviewed and corrected by the author. Document came from two books of [8, and 9]. The document uses Indonesian language and it is stored in a text file.

Searching relevant *sentences* against document uses Lucene (4.0 version) as search engine. Searching relevant *sentences* is performed by using *query* for *factoid* question and *non-factoid* question which is obtained in the *question analyzer* component. It searched to finding some *sentences* of the *document* that contains *query* for *factoid* or *non-factoid* questions. [10]

This section, *document retrieval* component will return 20 top ranking *sentences* based on the level of relevance to the *query* for *factoid* question or *non-factoid* question. This system uses *boolean query*, namely AND boolean operator in order to retrieve all *sentences* that have all words in *query*. [11]

3.3 Answer Extraction

This process is the final component of the question answering system to determine exact answer which is suitable to question which is entered by the user. The next After part of document is retrieved with *query* by Lucene [6, and 7]as search engine to obtain 20 relevant *sentences* in the *document retrieval* component, so 20 relevant *sentences* which are as answers candidate to question will be processed to obtain exact answer. To obtain exact answers, this system uses *expected answer type* and *clue word*. In this system, *answer extraction* is divided into two processes:

1. *Factoid* Question

In this component, the *output* which is obtained by *document retrieval* component will be *input* in *answer extraction* component. The *input* in this component is 20 relevant *sentences* which will be given score based on *query* occurrences and name entity tagging.

The steps for *factoid* question are:

- Every independence clause from *document retrieval* component is separated into token or word.
- Perform name entity tagging to each word which has entity.
- Every independence clause take all name entities that appropriate with the EAT as answers candidate. The answers of candidate that only contain *query* from question will not be included.
- Every answer candidate computed the distance between the answers candidate to all the *queries* from the sentence answer of question. Distance is calculated by counting the number of words between the answers candidate and *query.*
- Sort the answers candidate based on the distance between the answers candidate with *query* in the sentence.

2. Non-factoid Question

In this component, 20 relevant *sentences* which are obtained from *document retrieval* will be processed. For *non-factoid* question, *answer extraction* component uses pattern matching with sentence pattern based on [7] which includes all possibilities of *query* and *clue word* occurrences in the sentence. Sentence answer for each category of questions have patterns that are similar to each other. The difference lies in the *clue word* that follows the sentence answer for each question category.

The steps for *non-factoid* question are:

- The fist check pattern of sentence that is used in each sentence. It retrieved by *document retrieval* component. Then match the *query* occurrence in that sentence.
- The sentence with the highest find priority for each sentence. This sentence is chosen as answer candidate for question.
- The answers candidate based on the priority is sorted value in the sentence pattern.

The next step is making table of sentence pattern for *non-factoid* question and table of *clue word* based on [7].

3.4 Application Design

1.Use Case Diagram, 2. Use Case Narrative, 3. Activity Diagram, 4. Sequence Diagram

3.5 Graphical User Interface Design

This is the interface design of the system which is consists of header, question input box, search button, and answer box for answer which are provided by the system:

3.6 Construction

In this phase, the author constructs a system based on previous phases. The author will construct the system by using Java programming language and Netbeans platform. We used Lucene version 4.0 as search engine, and pattern matching method to obtain answer [10].

3.7 Graphical User Interface Implementation

In the implementation phase, Graphical User Interface (GUI) design which has been performed in workshop design is implemented on GUI implementation that can be seen similar with the Fig. 2 and Fig. 3.

4 Testing Scenario, System Testing and Result

System testing scenario is performed by attempting 30 questions that are arranged manually by the author. Those 30 questions are divided into 6 types of questions, there are: 5 questions for “Who” or “Who Is”, 5 questions for “Where” or “Where Is”, 5 questions for “When” or “When does”, 5 question for “What” or “What is”, 5 questions for “Why”, and 5 questions for “How” or “How does”.

The application made in Indonesian Language. This is system testing for factoid and non-factoid question types based on the system testing scenario which has been conducted by the author.

1. Factoid question

Figure 2. GUI Implementation for Factoid

2. Non-factoid question

Figure 3. GUI Implementation for Non-factoid

The assessment of the results refers to the system response [1]. Category the assessment of the results of the answer is:

Table 3. Category System Answers

Category	Notes
Correct	The system returns the correct answer and extracted from the relevant document
Not Agree	The system returns the correct answer but obtained from document that isn't relevant to the question
Not Correct	The system returns the answer in the form of string which is more than the actual answer or lose some of the answers and extracted from the relevant document to the question
False	The system returns the wrong answer.
No answer	The system cannot find an answer.

From the system testing which has been conducted by the author previously, this is the result of this experiment can be seen in the table below:

Table 4. The Test Results With The Passage 20

Category	TOTAL	Percentage
Correct	27	80%
Not Agree	1	0%
Not Correct	1	3.33%
False	1	3.33%
No answer	1	3.33%

Based on the table above, the author analyzes that from the scenario which has been conducted by the author previously, for 30 questions there are 27 question which have value "True" which are answer by the system, 0 question has value "Unsupported", 1 question has value "Not Correct", 1 question has value "Wrong", and 1 question has value "Not Answered". This case shows that this question answering system which has been created by the author is adequate work well because of 90% answer which is produced by this system provides the exact answer to a question.

5 Conclusion

Based on the discussions in the previous chapters the author concludes as follows:

1. The corpus which is used in this question answering system is taken from the previous research [8]. It has been corrected and reviewed by the author because the author found indication that the corpus still has weakness.
2. The architecture of question answering system that is described on chapter III is still not able to answer all questions that are entered by the user correctly. The results for each testing of the system give information that the system can answer 27 questions correctly is 90%, the question which has *"Not Correct"* category is 3.33%, the question which has *"not agree"* category is 0%, the question which has *"false"* category is 3.33%, and the question which has *"not answer"* category is 3.33%.
3. For *factoid* question, this question type has tested on system testing which is conducted by testing scenario and the resulting of testing for *factoid* question is quite good because all questions that are input by user can be answered by the system correctly.
4. For *non-factoid non-factoid* question, the resulting of testing process is not good enough because there are so many answers which aren't available (no answer) or not exact answers. Those answers are because the exact answers have more than a sentence pattern or usually there's no *clue word* which is suitable to question.
5. The best performance is shown by *Who*-type, *Where*-type, and *When*-type of question words because based on the testing result, the system is able to handle *factoid* question which is input by the user. Otherwise, for the unsatisfactory result is shown by *what*-type, *Why*-type, and *How*-type because the system is quite difficult to handle the answers which have more than one sentence pattern.

Based on the results of research that is conducted,We suggest the next researcher as follows:

1. The existence of further development of the system in identifying file type extensions such as: *pdf* and *chm* files that previously have not been identified by the system. To development of the next system can use all types of file extensions.
2. Using larger document so that the system has a larger database to answer the question that entered by the user.
3. For next Implementation expected to be developed within a wider scale implementation such as the application of internet.
4. Limitation of question words that can be used include: *"who", "when", "where", "what", "why", and "how"*; is still not cover all types of question words that will be able to answer all kinds of question words which are entered by the user.
5. The type of interrogative sentence that can be used in this system is still a single interrogative sentence, so the author suggests extending the interrogative sentence that can simultaneously handle two types of questions which is asked by the user.

Acknowledgment. This paper in conjuction with The 1st International Conference on Islam, Science and Technology (ICONIST 2018) in Malang, East Java.

References

[1] Naf'an, M.Z. *Sistem Tanya Jawab Berbahasa Indonesia Tentang Sejarah Khulafaur Rasyidin.* Jakarta: Department of Informatics Engineering Faculty of Science and Technology Syarif Hidayatullah State Islamic University (2012).

[2] Yusliani, N. and Purwarianti, A. *Sistem Question Answering Bahasa Indonesia Untuk Pertanyaan Non-factoid.* Bandung: School of Electrical Engineering and Informatics Bandung Institute of Technology (2011).

[3] Fukumoto, J. *Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method.* Japan: Ritsumeikan University (2007).

[4] Buettcher, S., Clarke, C.L.A., and Cormack, G.V. *Information Retrieval; Implementing and Evaluating Search Engines.* London: MIT Press (2010).

[5] Monz, C. *From Document Retrieval to Question Answering.* Netherlands: Universiteit van Amsterdam (2003).

[6] Lutfi, C.R. *Question Answering System on Juz Amma Translation Using Rule-based Methods*. Malang: Department of Informatics Engineering Faculty of Science and Technology Maulana Malik Ibrahim State Islamic University (2012).

[7] Zulen, A.A. and Purwarianti, A. *Study and Implementation of Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question*. Bandung: School of Electrical Engineering and Informatics Bandung Institute of Technology (2011).

[8] Al-Suyuti, I. *Tarikh Khulafa; Ensiklopedia Pemimpin Umat Islam dari Abu Bakar hingga Mutawakkil.* Bandung: Hikmah (2010).

[9] Yatim, B. *Peradaban Islam.* Bandung: Rajawali Pers (2010).

[10] McCandless, M., Hatcher, E. and Gospodnetic, O. *Lucene in Action (2nd Edition).* Shelter Island: Manning Publication (2010).

[11] Metzler, D., Croft, W.B. and Strohman, T. *Search Engines Information Retrieval in Practice*. Boston: Pearson (2010).