Indonesian Probabilistic Question Answering System for Enterprise Search of Hospital

Shinta Puspasari

STMIK MDP Palembang
shinta@stmik-mdp.net

Abstract: A probabilistic technique is proposed in this study to develop an Indonesian question answering system. The system processes factual type of question in natural language and finds the answer by retrieving information from documents. The system searches some answer candidates from a hospital text document for the input question about the hospital and presents an exact answer as the output. In terms of effectiveness, correctness of the answer is evaluated. Experimental result shows that the correctness is 60% in average. The QA system, which is part of next development of enterprise search system, can be made for supporting customer service, and increasing customer satisfaction.

Keywords: Question Answering, Probabilistic Technique, Enterprise Search, Information Retrieval

1 INTRODUCTION

Hospital is a public service enterprise. The organization aims to serve the people who need help in case of healthy. People come to the hospital to keep the healthy for their life. The customer will choose a hospital that can give best service by searching the information about its service to ensure that their life can be saved by the right hand. The accurate information, easy, and fast transfer of information to customer will affect the customer satisfaction and trust to the hospital. Many hospitals take the operator to give information to customer. Inaccurate and slow information transfer often happens because of human weakness. The operator needs to search the information from other people, or a stack of document that is time consuming. In emergency case, time becomes an expensive cost. Slow and inaccurate information will make fatal error. This problem will affect the customer trust and also the enterprise income.

Computer technology is the solution of that problem. Computer can search the information accurately and fast. An application system can be built to answer all questions from customer about the hospital. The system is called Question Answering System. Question Answering (QA) is different to search engine that returns a list of relevant documents. QA system retrieves an answer of question that relevant to user question in natural language. This system searches the answer of submitted question from a collection of textual document. Text processing technique is used for developing the system.

QA system has been developed using many techniques, i.e using syntactic and semantic analysis [1] to perform answer extraction, probabilistic approach [2] for phrase ranking, and statistical approach [3] that used support vector machine-based extractor, but it was implemented in different domain makes those techniques cannot be compared each other to find the best one. Developing QA system for specific domain is more difficult because it has specific word and phrase that is different in meaning to the common word and phrase [4]. QA system for different domain can be generated using the same technique, but may have the different performance. Especially for Indonesian QA system that using some developing techniques, which performance is not optimum, for Indonesian text processing.

This study tries to propose a probabilistic technique for developing QA system of specific domain, which is hospital enterprise. The document is modeled in bigram, and probabilistic approach is used for answer extraction. The system uses Lucene 1.9 for document retrieval generated in Java.
programming language. To test the performance of QA system, correctness of answer that present by the system is evaluated. In experiment, a collection of document from a public hospital is used as data set for answer retrieval. Factual question in Indonesian, which about the hospital, is submitted as input query. The question expects an answer about person, organization, date, and location. The result shows that the proposed technique is effective to answer the question. The system can be used for supporting customer service in order to increase customer trust and loyalty.

The rest of this paper is organized as follows: section 2 presents theoretical background of this study, section 3 illustrates the experiment in developing the application system, and the result of the proposed technique that concluded at the end of paper for future study.

2 THEORETICAL BACKGROUND

2.1 Overview Question Answering

Question-Answering (QA) is a process of retrieving the exact answer for question in natural language from a collection of document [5]. The question will be processed by using natural language processing technique to get the key word for searching set of answer candidates. The correctness of answer is used for system effectiveness evaluation [6].

2.2 Question Answering System Architecture

The architecture of QA system consist of some phases, i.e Query Modulation, Document Retrieval, Sentence Retrieval, Answer Extraction and Answer Ranking [2] that illustrates in the following Figure1.

a. Query Modulation

The process of converting question in natural language into query that is used for document retrieval. The question will be identified grammatically and split into words in order to find the key words. The query contains key words of question for searching relevant document.
e. Answer Ranking

Ranking the answer candidates and taking the top-1 candidate as the exact answer, which has the highest score.

2.3 Probabilistic Approach for Natural Language Processing

The probabilistic of retrieving term $t$ in document $d$ can be calculated using Likelihood estimation that formulize as the following eq.1 [7],

$$P_{mle}(t \mid d) = \frac{tf_{t,d}}{N_d} \quad \cdots (1)$$

where $tf_{t,d}$ is the frequency of term $t$ in $d$, and $N_d$ is the total number of words in document.

2.3.1 Bayesian Rule

Bayesian rule is used to calculate the probabilistic of retrieving two words at the same time (bigram). The following eq.2 illustrates the rule [8]. The equation shows that the probabilistic of $B$ is dependent to $A$,

$$P(A \mid B) = \frac{P(B \cap A)}{P(A)} \quad \cdots (2)$$

2.3.2 Bigram Linear Interpolation

Interpolation is processed to bigram model of the document for smoothing the estimation result of probabilistic bigram that caused by unigram data sparseness [8]. The basic interpolation is liner interpolation. This interpolation can be used to estimate the probabilistic of term $t_{i-1}$ and $t_i$ appearance in document $d$. The linear interpolation function is [7],

$$P(t_{i-1}, t_i \mid d) = \lambda_1 \times P(t_i \mid d) + \lambda_2 \times P(t_{i-1} \mid d) \quad \cdots (3)$$

where $\lambda_1 + \lambda_2 = 1$ is two parameters whose value must be tuned based on the training data.

2.3.3 Hidden Markov Model

Let $X= (x_1, x_2, \ldots, x_T)$ is a sequence of random variable taking value in some finite set $S = (s_1, s_2, \ldots, s_N)$ as space state. The Markov model has the following eq.4 property [8],

$$P(x_{T+1} = s_k \mid x_1, x_2, \ldots, x_T) = P(x_{T+1} = s_k \mid x_T) \quad \cdots (4)$$

In hidden Markov model (HMM), the sequence of states that will be passed by the model are unknown. Only its probabilistic function is available.

3 EXPERIMENT

The experiment uses a data set of document in Indonesian that contain 18 text documents from web page of a public hospital in Jakarta and a document from department of healthy that are collected automatically using a crawler. The total amount of data set is 136 Kb.

The system receives factual type question which ask person, object, place, date, and availability. The question must be typed grammatically right in Indonesian and ask anything about the hospital.

For document retrieval, the QA system uses the available information retrieval system, which is open source, Lucene 1.9. Lucene index and searches the relevant document based the input question. This system return top-N relevant document to the submitted query. In implementation, QA system is developed by using Java programming language.

3.1 Document Preprocessing

Firstly, document is preprocessed using some method in some steps. The preprocessing phases are illustrated in the following Figure 2.
Noise removal aims to remove noise from the document, such as html tag, that may affect the model formulation. Tagging gives name entity to entity word in document, such as person, date, organization, and location. The tagged word will be used for answer candidate searching process.

The training document is used for tuning the parameter to find the optimum value, and the bigram model of each document is generated.

### 3.2 System Architecture Design

The QA system is generated in Java language, and the architecture design is represented by Figure 3. User submits a question to the system that treats it as query. Based on the query, Lucene search the relevant document and retrieve top-5 documents that will be extracted to find the answer candidates.

Suffix and prefix are not removed (stemming process), because the experiment focus on proposed technique, that makes the searching method is exact match. It means that the system will search the same word in document as the given query based on the question. Testing queries are chosen based on the word in document collection. The query is analyzed to find the key word for sentence extraction and reduce the document into passage which may contain answer candidates. After document retrieval, next step is passage retrieval. Named entity tag in document is used for scoring and ranking the passage. Top-1 passage of each document, which contains 2 sentences, is retrieved. The passage will be extracted by answer extractor, which is generated using Markov model, and produce some answer candidates. For bigram linear interpolation, the experiment takes parameters $\lambda_1 = 0.1$ and $\lambda_2 = 0.9$. The values are taken as the result of tuning process using training data. Scoring and ranking to the answer candidates base on probabilistic estimated value of language model of training document. The highest rank answer candidate will present as the exact answer to user.

The system are tested using four type of question, i.e asking availability, person or organization, date, and location related to the hospital. 5 (five) queries are given for each type of question, and two of the query are NIL question, which is no answered question. For this type of question, the answer is correct if NIL is presented as the answer. Table 1 represents some questions in natural language that submits to the system for testing.

To extract the answer for the question that asks availability of an object, the technique process bigram model of document without deep search using Markov model. This technique decrease computational time for searching.

### Table 1: Testing Queries

<table>
<thead>
<tr>
<th>Type of Question</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>Adakah klinik mata ?</td>
</tr>
<tr>
<td>PERSON or ORGANIZATION</td>
<td>Siapa Alexander ?</td>
</tr>
<tr>
<td>LOCATION</td>
<td>Dimana alamat ?</td>
</tr>
<tr>
<td>DATE</td>
<td>Kapan diresmikan ?</td>
</tr>
</tbody>
</table>

### 3.3 Result and Analysis
Some constrain exist during the experiment. Many phrases in document need to be identified firstly, for example “dr” which means doctor. Miss identified phrase makes mistake in modeling the document. Some words has high frequency but cannot be eliminate because will effect the identification of answer candidate. The low performance of tagger generated for Indonesian document also makes miss identification to the answer candidate. The tagger often fails to give the correct name to the entity in document, or gives the wrong name to the entity. The mentioned problem affects the generated QA system performance.

Scoring technique for passage retrieval can retrieve the correct passage contains 2 (two) sentences that will be extracted based on the named entity in it. This entity represents the answer candidate of given question. The experimental result is illustrated in the following Figure 4 that shows QA system screenshot,

Figure 4 (a) (b): QA Application System Screenshot

Effectiveness of the system is evaluated using correctness indicator. User gives the feedback to evaluate the answer whether it is correct or incorrect.

Figure 5 shows that presentation of correctness of the answer presented by QA system is low, which is 60% in average. This happen especially to the query that consist of more than 2 words (including question word). In this case, probably the extractor fails to find the exact answer because it is affected by probability distribution of each answer candidate. The longer query makes the increment of state that may be through by HMM. The answer candidates will have small and uniform probability that makes difficulty in finding the correct answer.

The number of state that may be through by HMM makes the system consuming more time to calculate the probability of answer candidate. Searching process may take computation cost as \( O(n^t) \), where \( t \) is the number of state.

Figure 5: Percentage of Correctness

4 CONCLUSIONS AND FUTURE WORK

The proposed technique is effectives to retrieve relevant answer, which is extracted from a set of hospital document, to the question. Probabilistic approach for developing the system can be used for developing QA system, which is built as part of hospital enterprise search system. The experimental result performs 60% of correctness in average when was tested using 15 factual questions about the hospital. What, where, and when submitted question have low answer correctness.

This happen because the experiment have some problem and limitation. The document is processed using the low performance tagger for Indonesian. Many entities in document, i.e person, organization, date, and location, fail to be tagged by the tagger. This affect the performance of sentence retrieval and answer extractor to retrieved the exact answer. Stemming may affect the system. The QA
system does not stem the document and answer. This makes the system cannot retrieve relevant document that has no query key word in it.

The proposed technique need to be improved for increasing the QA system performance. Stemming may improve the performance because it can make the system retrieve relevant document that contains no key word but has the same context. The system can support customer service for increasing customer satisfaction.

REFERENCES


