Naive Bayes Classification in The Question and Answering System

Viny Christanti Mawardi [1)] Jeanny Pragantha [2)] Sakti Sarjono [3)]
[1) 2) 3)] Department of Computer Science
Tarumanagara University
Jakarta, Indonesia
email: viny@untar.ac.id [1)] anlibra25@gmail.com [3)]

***Abstract*—Question and answering (QA) system is a system to answer question based on collections of unstructured text or in the form of human language. In general, QA system consists of four stages, i.e. question analysis, documents selection, passage retrieval and answer extraction. In this study we added two processes i.e. classifying documents and classifying passage. We use Naïve Bayes for classification, Dynamic Passage Partitioning for finding answer and Lucene for document selection. The experiment was done using 100 questions from 3000 documents related to the disease and the results were compared with a system that does not use the classification process. From the test results, the system works best with the use of 10 of the most relevant documents, 5 passage with the highest score and 10 answer the closest distance. Mean Reciprocal Rank (MMR) value for QA system with classification is 0.41960 which is 4.9% better than MRR value for QA system without classification.**

Keywords—Naive Bayes Classification; Dynamic Passage Partitioning; Question Answering; Information Retrieval.

I. INTRODUCTION

Nowadays there are many websites that provide health consultation in the internet, such as http://www.konsultasikesehatan.com and http://www.tanyadokter.com. Sometimes, answers can be obtained within a few days. One of the reason is lack of professionals skilled who can answer that question. The situation can be different if the website implement the Question and Answering System, so the answer can be found automatically.

In search of information, the Question and Answering System use time efficiently. However, similar studies are still rare for the Indonesian language documents although the population of Indonesia is the fourth largest in the world. According to the Ministry of Communications and Information Technology, Internet users in Indonesia have reached 45 million people [1]. With such big population, Question and answering systems in Indonesian documents are indispensable.

In previous work, we used passage-based retrieval to find the answer, Stanford NER for named entity recognition, Lemur to retrieve relevant document and Indonesian legal document for source document of QA system. This QA system produces 72% accuracy on the retrieval of top-5 documents. The disadvantages of this research was the retrieval of relevant documents were not on the top-5 documents and distribution of passage is improper so some information were lost.

In this research, we used classification concept to cluster documents. We used Naïve Bayes classification because it is an efficient and effective algorithm. Naïve Bayes has a competitive performance in classification area because of assumption of conditional independence is rarely true in real world applications [3]. Classification was used to filter the document so that the answer can be obtained from a set of documents which have the same topic, i.e. documents in the same cluster. Searching will focus on the collection of documents that were in one cluster. This will facilitate the process of finding and collecting candidate answers. The classification was performed to obtain passage on passage retrieval step.

In this research, we used Dynamic Partitioning Passage or Variable Length Arbitrary Passage Partitioning is used to get the answers. Dynamic Partitioning Passage is one of answer selection method that built a passage overlap with the previous passage. The purpose to construct overlap passage is to avoid separation of the relevant text. Stanford NER still used as a tool to recognize an entity. In the health domain, we used three entities, namely disease, drug and dosage. The document selection process is done by using Vector space models and Boolean model from Lucene (see: http://lucene.apache.org).

II. QA SYSTEM

The purpose of QA system is to identify the answer to the question of an unstructured collection of text in the form of human language [4]. The simplest form of QA system is factoid question. Factoid questions are questions that seek answers that are simple facts that can be found in short words [5]. Factoid question is usually asked questions about specific answers and easily categorized as the person's name, organization's name, and location's name.

Based on the scope of domain, QA system can be divided into Open domain and Close domain. Open domain QA is a system that can answer questions in any field. While Close domain QA is a system on specific areas such as fields of health, law and others. Type of QA system that was investigated in this study is close QA domain, which is health.

In general, the QA system is divided into four stages. The first stage is the question analysis stage, the second stage is the selection of documents, the third stage is the search Passage and the last stage is the extraction of the answer [4]. In this study, we will see the influence of classification to find the answers. We added classification in stage two and tree. In stage two, classification is used to select documents. In stage tree, the classification is performed on each passage. The QA systems block diagram can be seen in Figure 1.

In the first stage, the question will be analyzed to determine the types of answers. Analysis is conducted using several regular expressions which are developed based on Indonesian's grammar. The result is keywords derived from the question.

In the second stage, the documents were selected using vector space model and Boolean model. Keywords that are obtained from previous stage will be used to perform queries on a selected documents.

In the third stage, relevant documents are broken up into sections or passage using dynamic passage partitioning method. Furthermore, all of passage will be classified into several classes using the Naive Bayes method. Passage that has the same class as the query is ranked according to number of features that found in the passage. Passage that has a high ranked is forwarded to the next stage.

In the last stage, answer is derived from passage documents that is obtained from the previous stage. Answer candidate is extracted using the entity name recognition. Entity which has the same type with the answer is considered as answer candidate.

The QA systems were evaluated by calculating the Mean Reciprocal Rank (MRR). The evaluation is done by comparing the answers generated by the system with the answers provided by the researchers.

A. Question Analysis

The question analysis is done by detecting the type of questions to determine the type of answers. In this study, there are three types of answers that is disease, drug's name and dosage of medication. The analysis is done by the introduction of the question asked and the words it used. In Table 1, there is a relationship between the word used by a question and its answers.

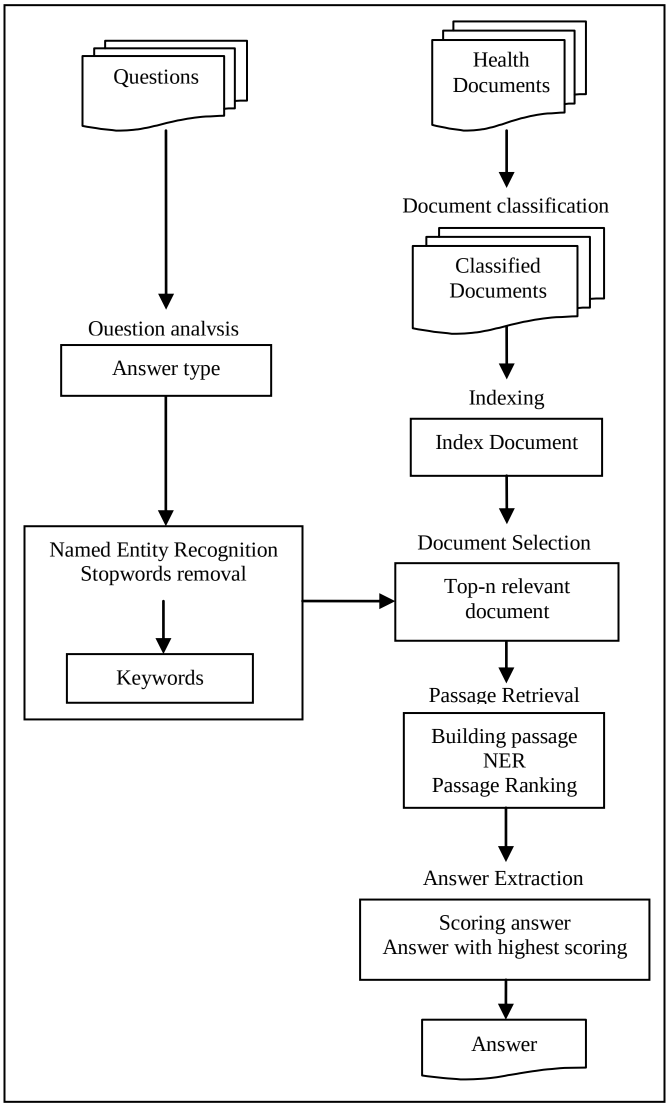

Figure 1. QA systems for Indonesian health document.

TABLE I. RELATIONSHIP BETWEEN THE TYPE OF QUESTION WORDS AND THEIR ANSWERS

Types of answers	Question words
Disease	• *Apakah Penyakit ..?* • *Apa Penyakit ..?* • *Penyakit apakah ..?* • *Penyakit apa ..?* • *Penyakit Manakah ..?*
Drug name	• *Obat Apakah ..?* • *Apa Obat ..?* • *Obat Apa .. ?* • *Obat Manakah .. ?*
Dosage	• *Berapakah dosis .. ?* • *Berapa dosis .. ?*

One way to identify the type of the answer is to give the word an entity using the Named Entity Recognition (NER). It is a process to extract information to classify entities into predefined categories [5]. In this study, NER is used to identify the entities associated with the answers. In general NER focuses on classifying categories such as names of people, locations and organizations [5]. In this research we classify entities into three types, i.e. disease, drug's name and dosage of the drug.

NER can be considered as a process of classification so that the methods used for classification can be applied to the entity name recognition. Examples of methods that can be used for perform recognition is Naive Bayes, Hidden Markov Models and Conditional Random Fields. Some of the leading NER has been made and used widely, such as Stanford NER, Lingpipe, and GATE. In this research we used Stanford NER to identify the entities associated with the answers. Stanford NER is one of the NER recognition system made by the Board of Trustees of The Leland Stanford Junior University [6].

B. *Document Selection*

The selection of documents is a process that is done before searching the answers. So the answers will be searched from a collection of documents which are relevant to the query. In this study we use Vector space model and Boolean model.

In Vector Space Model, each document and query is represented by vector. Furthermore, a method must be chosen to calculate document similarity with the query. The similarity of two vectors can be seen from different points. Boolean model is one of the strategies used in Information Retrieval to obtain documents relevant to the query [5]. As the name implies, this model determines the relevance of document with the operation of Boolean algebra using logic operators 'NOT', 'OR', or 'AND' in queries.

C. *Dynamic Passage Partiioning*

Dynamic Passage Partitioning or Variable Length Arbitrary Passage Partitioning is used to partition a document into smaller parts to improve the accuracy in the retrieval process. Dynamic Partitioning Passage will search the document to find the same word as the query then build a passage around that word. If the word is found in position n, then the passage is built starting at position n and so on until position n + p with p is the size of the passage. Subsequent passage built overlap with the previous passage and begins in position n + (p/2). The purpose of the construction of passage overlap is to avoid separation of the relevant text. The size of the passage is between fifty and six hundred words.

The use of Dynamic Partitioning Passage in this study because, this method has higher accuracy in passage-based retrieval than other methods. Marcin Kaszkiel mentioned that, the optimal result was obtained from the query of more than ten words and the passage of two hundred words. As for the query that is less than ten words, the optimal results will be obtained from the passage of two hundred and fifty words [9].

D. *Naive Bayes Classifier*

Naïve Bayes classifier is a probabilistic classifier to apply Naive Bayes theorem [7]. Naive Bayes classifier assumes that the presence or absence of a particular feature of a class is not related to the presence or absence of other features [8].

Class formations are conducted by researchers by reading some of the articles as a sample and proceed with the classification which is done manually by a researcher to classify documents into some of the most dominant class. In the Naive Bayes classifier, the probability formula to classify is:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \qquad (1)$$

In text classification, its main purpose is to find the most appropriate class for the document [8]. The most appropriate class in the Naive Bayes classifier is a class that has a maximum or a posteriori probability greatest of all classes:

$$c_{map} = \max_{c \in C} \hat{P}(c|d) = \max_{c \in C} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c) \qquad (2)$$

P is changed to $\hat{P}$ because the formula above the original value of P (c) is not known, but estimated from the training data [8]. Training conducted to estimate the value of the parameter P (c) and P (t | c) in the following formula:

$$\hat{P}(c) = \frac{N_c}{N} \qquad (3)$$

Where N_c is number of document in class c and N is number of document in training.

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \qquad (4)$$

Where T_{ct} is word frequency in document of class c and V is the vocabulary. To avoid the possibility of the emergence of the word with probability 0 which caused by the absence of a word in a class, Laplace smoothing is used.

In practice, Naive Bayes method should be trained to build a classifier. The training is done in the following steps:

1) *Vocabulary formation of training data*

2) *Perform calculation for each class with formula number 3*

3) *Calculatee the conditional probability of each word in each class using formula number 4*

After classifier is formed, then use the article with the following steps:

1) *Make a list of words from the test article that appeared in the vocabulary*

2) Calculate the probability for each class article using the formula number 1

3) Determinate the class for the test article using the formula number 2

E. Passage Ranking

In QA, candidate answer is obtained from the passage that has a high rank. In general, the first step to rank a passage is done by Named Entity Recognition or classification of answer's type on all passage. Type of the answers obtained from a question will provide an information to discard the passage that is not contain the answers. The passages remain then to be ranked with the features made by human and by machine learning techniques with supervised training [5]. Examples of features commonly used to rank the passage [5]:

1) The number of entities in the corresponding passage.

2) The number of keywords contained in the question passage.

3) The longest word count of keyword questions according to the question.

4) Ranking of document in which the passage partitioned.

F. Answer Extraction

Answer Extraction is the last stage in the Question Answering. In this stage, answer is estraed brom the high rank passage. The answers for factoid questions can be done by extracting all the entities that are candidates the answers [5].

The process of extraction the answer is as follows:

1) Calculate the sum of the distance between candidate answers and each keyword.

2) Candidates answer that has more keywords are prioritized.

3) If there are several pieces of candidate answers have the same distance, then the answers selection is done by selecting the candidate answers which has higher document rank or higher passage rank.

4) If the candidate answer has the same ranked document or passage, then the candidate answer is the answer with the higher frequency from the list of candidates' answers.

5) If still equal, then the selection of the candidate's answers conducted from the first answer was processed.

G. Mean Reciprocal Rank

There are many techniques that can be used in evaluating a QA system. The most commonly used technique is the technique of the Mean Reciprocal Rank (MRR) which is a technique provided by TREC in 1999 [5]. Before evaluate, the system is provided a set of questions and answers. MRR evaluate the system against a set of questions and answers to that question. MRR further compare the resulting answers with the answers provided. Each question is rated according to the ranking of the correct answer. The equation to calculate the MRR is as follows:

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N} \qquad (5)$$

III. EXPERIMENT

Number of articles used for this study are 3000 articles. Articles are obtained from several sources i.e. kompas.com, tanyadokter.com, drug-penyakit.com, cariobat.blogspot.com, and medicastore.com. The experiments were performed using the 100 questions that have been prepared. Testing is done with 3 schemes that have the specified number of the most relevant documents, the number of passage with the highest score and the number of answer with the closest distance. Scheme of experiment can be seen in table 2.

TABLE II. SCHEMES OF EXPERIMENT

Scheme	Experiment
Scheme 1	QA System with passage classification
Scheme 2	QA System with document classification
Scheme 3	QA System without classification

Each scheme was divided into 3 stages i.e., first stage to determine the number of relevant documents that have the best results, the second stage to determine the number of passage that has the best results and the final stage the number of answers that have the best results. Table 3 show numbers of document, passage and answer.

TABLE III. STAGE AND NUMBERS OF DOCUMENT, PASSAGE AND ANSWER

Stage	Experiment 1	Experiment 2	Experiment 3
Stage 1	Top-5 Doc, Top-5 Passage, Top-5 Answer	Top-10 Doc, Top-5 Passage, Top-5 Answer	Top-15 Doc, Top-5 Passage, Top-5 Answer
Stage 2	Top-10 Doc, Top-5 Passage, Top-5 Answer	Top-10 Doc, Top-10 Passage, Top-5 Answer	Top-10 Doc, Top-15 Passage, Top-5 Answer
Stage 3	Top-10 Doc, Top-5 Passage, Top-5 Answer	Top-10 Doc, Top-5 Passage, Top-10 Answer	Top-10 Doc, Top-5 Passage, Top-15 Answer

The formation passage performed by splitting the top-n documents obtained from the search of relevant documents. In this study, the size of the passage is determined with the following requirements: 200 words to query more than 10 words and 350 words for the queries that are less than 10 words. Determining the size of the passage is based on research results Marcin Kaszkiel its consideration of research on document collections used in this study [9]. Passage Ranking is done by weighting its features, see table 4.

TABLE IV. VALUE OF PASSAGE RANKING

Value	Feature
5.0	Passage contains the same entity with the types of answers
0.5	Passage contains more than one entity to the same type of answers
1.0	For passage, which is part of the top-5 documents
1.0	For the passage containing the keyword more than half the number of keywords
1.0	To passage the word sequence and in accordance with the keywords at least half of its length keyword

IV. RESULT

Experiment for Naive Bayes classifier done by providing 100 test articles. Test article consists of 50 articles for each class. Articles used for testing have the same form as well as a collection of articles in the training data. Evaluation is done automatically by the program. In a 100 test articles, Classifier can correctly classifying all articles which 100% accuracy. These results are due to a very good training articles and clarity between the two classes that are classified.

Name entity recognition (NER) test is done by providing 2299 words of 10 articles testing. Evaluation for NER is done manually in which the researchers read the results of the entity's name and count the number of words correctly recognized. The test result shows that from 2299 words, NER failed to classify 76 words consisting mostly words of the entity disease's name and the medication dosage. The Classifier can classified all articles correctly, in which reach 100% accuracy.

From the experiment we get the best result in top-10 document and top-5 passage. Table 5 shows the result of experiment from 3 schemes with top-10 document and top-5 passage in 3 different top-n answers. From 3 kinds of top-n answers we can see the best result is scheme with top-10 document, top-5 passage and top-10 answer.

In table 5, we also can see that the best result is experiment in scheme 3 with MRR 0.427. Scheme 3 is experiment QA system without classification. But when we see the scheme 1 and 2, the best result is QA system with document classification which is scheme 1. The result QA system with classification have smaller MRR compare QA system without classification but not significant.

TABLE V. TABLE STYLES

Scheme	Top-10 Doc, Top-5 Passage, Top-5 Answer	Top-10 Doc, Top-5 Passage, Top-10 Answer	Top-10 Doc, Top-5 Passage, Top-15 Answer
1	0.41050	0.41960	0.41960
2	0.39566	0.40739	0.40739
3	0.41716	0.42738	0.42738

V. DISCUSSION

Misclassification due to drug doses writing as an example of inconsistency. Such as, the inclusion of frequency of drug use in a day ("*3 gr dalam sehari*") and how to use ("*secara oral*"). Classification errors that occur in the name of the disease are excess recognition. For example, the word "*gangguan pembuluh darah balik penyumbatan*" that are recognized as overall disease which in fact only "*gangguan pembuluh darah*".

At first, the three schemes of experiments aimed to determine the effect of classification in QA system. Testing and experiments shows that the scheme no 3 which does not use the classification has a better MRR than two other schemes. This happens for several reasons, namely: first, an error occurred while classifying questions. Although the testing classifier get 100% accuracy, but at the time there were still classifying errors due to the small size of question (about 5 to 20 words).

Naive Bayes classifier is used to classify the word identity. But when the number of words is too small and there is no recognizable words, then Naive Bayes is failed to classify the word. This failures made experiments on the scheme 1 and scheme 2 could not find the answer. The other reason is, threre is a shift in of correct answers to a lower rating. This happens because after classifying, several articles and passage that produces the correct answer has a smaller value than the actual answer. This small value makes the rank of documents and passage which have the correct answers to be down.

But there is also the question successfully answered by scheme 1 and 2 but cannot be answered by the third schemes namely, the question no 30. Table 6 shows that the answer "*hiv*" not found in the top-5 answers for QA system without classification (scheme 3). This is caused by the classification process function as filter for the article and the passage so that only articles that are really relevant returned by search engines. Therefore, the answers filtered by the classification process and the actual answer into the top-n answer.

TABLE VI. ANSWER FROM 3 SCHEMES FOR QUESTIONS NO 30

Question no 30 is: "*penyakit apa yang paling berisiko mengenai seseorang apabila sering melakukan seks bebas?*"			
The real answer is: "HIV"			
Ranking of answer	Scheme 1	Scheme 2	Scheme 3
1	*hiv*	*hiv.*	*ed.*
2	*null.*	*alergi.*	*penyakit*
3	*null*	*flu.*	*jantung.*
4	*null.*	*edema.*	*depresi.*
5	*null.*	*asma.*	*impotensi.*
			osteoporosis.

VI. Conclusion

The experiment with 3 types give the best result where QA system in top-10 document, top-5 passage and top-10 answer. The best MRR in classification passage for QA system (scheme 1) is 0.41960. The best MRR in classification document for QA system (scheme 2) is 0.40739. And the best MRR in QA system without classification (scheme 3) is 0.42738.

Although QA system without generating classification has MRR values slightly higher, we managed to prove that the concept of the classification can increase the rank of real answers. Documents or answers filtered by category to make the answers can be found in the top-5 answers. The concepts of classification successfully filter out the document and increase the answers to the QA system.

In further, we must improve the classification process in order to produce a cluster of document which is more appropriate to answer question. Additionally passage dynamic partitioning method can be developed further into more specific QA system by changing the candidate answers as a basis for the distribution of passage.

References

[1] Wahono, Tri(2010), “Pengguna Internet di Indonesia capai 45 Juta”, http://tekno.kompas.com/read/2010/09/20/15412739/Pengguna.Internet.di.Indonesia.Capai.45.Juta-12.

[2] M., Viny Christanti, Non Vie and P. Jeanny. “Question Answering System for Indonesian Legal Documents”. In Proceeding of Quality in Research 2011, vol. 1, p: B14.3, 190, July 2011.

[3] Zhang, Harry, “The Optimality of Naïve Bayes,” in Proceedings of the 17th Florida Artificial Intelligence Research Society Conference, 2004.

[4] Chen, Jiangping; Diekema, Anne R.; Taffet, Mary D.; McCracken, Nancy; Ozgencil, Necati Ercan; Yilmazel, Ozgur; and Liddy, Elizabeth D., "Question Answering: CNLP at the TREC-10 Question Answering Track". Center for Natural Language Processing. Paper 2, 2000.

[5] Jurafsky, Daniel and Martin, James H., “Speech and Language Processing: An Introduction to Natural Langeage Processing”, Computational Linguistics and Speech Recognition, Upper Saddle River : Prentice Hall, 2009.

[6] The Stanford Natural Language Processing Group, “Stanford Named Entity Recognizer”, http://nlp.stanford.edu/software/CRF-NER.shtml, 2011.

[7] Flach, Peter A. And Lachiche, “Nicolas. Machine Learning”, Boston : Kluwer Academic Publisher, 2003.

[8] Manning, Christoper D.; Raghavan, Prabhakar and Schutze, Hinrich, “An Introduction to Information Retrieval”. Cambridge: Cambridge University Press, 2009.

[9] Kaszkiel, Marcin and Zobel, Justin, “Effective Ranking with Arbitary Passages”, Journal of the American Society for Information Science and Technology, Vol. 52, No. 4. (22 January 2001), pp. 344-364