ISBN: 978-1-948012-05-8

Contents List available at VOLKSON PRESS

Intelligent Computing and Information Engineering (ICIE)

DOI : http://doi.org/10.26480/icie.01.2017.60.66

Journal Homepage: : https://www.intelcomp-design.com/

INTELLIGENT QUESTION ANSWERING METHOD FOR ENTERPRISE LAW

Zheng Xiaosong, Rao Wenbi

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, People's Republic of China.
**Corresponding Author email: {zxsong, wbrao }@whut.edu.cn*

ARTICLE DETAILS

Article History:

Received 12 May2017
Accepted 12 July 2017
Available online 14 September 2017

Keywords:

intelligent answering system, enterprise law field, question classification, answer extraction.

ABSTRACT

This study started from the special demand of enterprises, and discussed key technologies and methods in enterprise law field, aiming at classifying the problem of intelligent question answering system. In this paper, an enterprise law question set was established, and SVM method was utilized to realize question classification in enterprise law field, with classification accuracy as high as 84%. A multi-strategy fusion based answer extraction algorithm was proposed by fusing multiple answer extraction strategies, including keywords, semantics, and bag of words (BoW), and the answer extraction accuracy achieved up to 73%.

1. Introduction

With the development of market economy and increasing improvement of enterprise law, rapidly and effectively obtaining legal knowledge closely related to the interests of enterprises has become their urgent need. The information returned by traditional search engines is redundant, error, and irrelevant, which cannot satisfy legal consultants' demand for obtaining accurate answers. Therefore, intelligent answering system in enterprise law field emerges as the times require [1].

Traditional question answering system mainly includes three parts, namely, question classification, information retrieval, answer extraction. Since question answering systems oriented to enterprise law field mainly search answers from law text according to users' questions, without need to search relevant information from network resources, question classification and answer extraction were studied.

Question classification methods are mainly categorized into two classes, namely rule based methods and statistics based methods [2,3]. In the study on English text classification, a group of scientists extracted the bag-of-words (BoW) features and the bag-of-ngrams (bon) features of questions, and adopted a support vector machine (SVM) classification algorithm based on tree kernel, which achieved classification accuracy of 90.0% in 6 major classes of questions [4]. In other study, a group of researchers proposed a sparse network model (SNoW) based machine learning classification method, and established a semantic layer classifier, with classification accurate ratio of 88% [5]. It can be concluded from existing question classification methods that these methods include SNow, CNN, SVM, naive Bayesian methods, and the SVM based classification method was adopted in this study.

As the final step of an answering system, answer extraction is the key of the success of the system [6,7]. A group of researchers also proposed a pattern learning based method when studying answer extraction, which retrieves the answers of users' questions by constructing answer generation models. Experimental results have demonstrated that using pattern extraction answer extraction algorithm is conducive to improving the performance of OA systems [8]. For the questions and answers of natural language in open domain, a group of them also proposed noumenon based answer extraction algorithm, which calculates the similarity between a question and candidate answers to obtain the answer to a question by extracting the morphology, syntax and other features of the question and candidate answers [9]. From the study on answer extraction, it can be concluded that the mainly used answer extraction algorithm is based on sentence similarity. In this study, on the basis of traditional sentence similarity calculation methods, an answer extraction method based on multi-strategy fusion was proposed.

2. QUESTION CLASSIFICATION

Question classification mainly consists of three steps, including determination of question categories, constructing question datasets, and determination of classification methods.

2.1. Question categories

Question categories are the basis of choosing answer extraction strategy, and a reasonable set of question categories is of great importance for a question answering system. In this study, the six major Chinese question categories proposed by Harbin Institute of Technology were referred, and questions in enterprise law field were divided into five major categories, i.e., human category, number category, description category, enumeration category and definition category.

2.2. Question data set

Question data set is a kind of inevitable resource in statistics based classification [10,11]. In China Legal Counseling Center (http://www.zgflzxzx.org/) and Find Law Net (http://china.findlaw.cn/), a total of 1038 enterprise law questions were collected, and duplicate removal and screening were performed for

them, establishing a question dataset in enterprise law field, ELQ_set, containing 873 questions. The distribution status of the question dataset is shown in Figure 1.

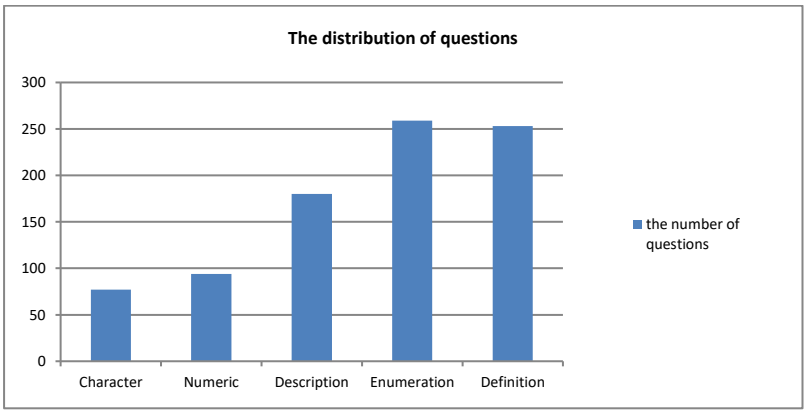

Figure 1: Distribution of enterprise law question instances.

2.3. Question classification method

Commonly used question classification methods are rule based and statistics based. Since Chinese grammatical rules are numerous, it is difficult to enumerate them by manual working, and therefore the performance of rule based Chinese question classification is unsatisfactory. For statistics based methods, since question sets in enterprise law field for beta are lacked, and SVM can still obtain relative optimal results on self-established question sets with good generalization ability, SVM based classification method was adopted in this study. SVM classification method is a statistical theory based classification method, with algorithm description as listed in Algorithm 1.

Algorithm 1 SVM multi-classification algorithm

Input: Question dataset $D=\{(\mathbf{x_1},y_1),(\mathbf{x_2},y_2),\ldots,(\mathbf{x_n},y_n)\}$, where $\mathbf{x_i}$ are sample data, or the feature vectors of questions, and y_i are the classification labels of sample data {1, 2, 3, 4, 5}, $i=1,2,\ldots,n$, and n is the number of training samples;

Output: SVM model with determined parameter set $C=(H,S)$, where H is the parameter of kernal function, and S is multi-classification model.

Begin

Divide D into k equal parts, i.e., T_1, T_2, ..., T_k, and take one part as the test set Te each time, with the rest for training set Tr

For i=1 to k

Te=Ti, Tr=D-Te

//Choose samples in training set and use SVM for training, resulting in C_i

$SVM(Tr)$, resulting in parameter set $C_i=(H_i,S_i)$

Set the kernel parameter of SVM as H_i, multi-classification model as S_i, resulting in new SVM_i

//Input test samples into newly constructed SVM_i model, obtaining accurate ratio of classification results P_i

$SVM_i(Te)$, obtaining P_i

Save (C_i,P_i)

For End

Choose the parameter set $C=(H,S)$ with highest accurate P_i as the optimal parameter set output, determining SVM model.

End

3. ANSWER EXTRACTION

According to question classification results, different methods are used in the answer extraction part to obtain answers. For complicated categories of questions, including enumeration, description, and definition categories, answers are usually a big paragraph of text. Therefore, usually, the similarity between the user input question and the paragraphs in the paragraph sets of candidate answers are calculated, resulting in the paragraph with highest similarity degree for returning to users. For simple categories like human and number categories, answers are generally a word or a phrase, and answers can be directly found by named entity recognition. Since named entity recognition technology is relatively mature, no much demonstration is made, and answer extraction for complicated questions (description, enumeration, and definition categories) are studied.

The answer extraction of complicated questions is actually to calculate the similarity between questions and candidate answers in candidate answer set. Currently, sentence similarity calculation is considered in aspects of keywords, BoW features, semantics features, etc. The three aspects of features were fused in this study, and a multi-strategy fusion based answer extraction algorithms was proposed.

3.1. Multi-strategy fusion based answer extraction algorithm

From the current study situation, it is clear that the answer extraction accuracy of single features is not high. The reasons are: ① keyword and BoW based answer extraction algorithms only consider words, ignoring word meanings; ② semantics based answer extraction algorithms only consider word meanings and ignore the essential usage of words. To deal these problems, a multi-strategy fusion based answer extraction algorithm combining multiple answer extraction strategy, including keywords, semantics, and BoW, was proposed, which considers similarity between questions and candidate answers in multiple layers, and more completely calculates the similarity between sentences, finally presenting reasonable and accurate answers to users. The definition of the sentence similarity between question Q and answer W is given by:

$$allSim(Q,W)=\alpha_1\times keySim(Q,W)+\alpha_2\times wsSim(Q,W)+\alpha_3\times vsmSim(Q,W) \quad (3\text{-}1)$$

Where

- $keySim(Q,W)$ represents the sentence similarity between Q and W based on keywords;
- $wsSim(Q,W)$ represents the sentence similarity between Q and W based on semantics;
- $vsmSim(Q,W)$ represents the sentence similarity between Q and W based on BoW;
- α_1, α_2 and α_3 represent the weights of the three basic algorithms in fusing multiple answer extraction strategies, which are determined by the answer extraction accuracy of them, and satisfy $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_3 > 0$, $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

3.1.1. Sentence similarity method based on keywords

For keywords, the similarity between two sentences is determined by the number of identical keywords, distance similarity between keywords, and the order similarity of keywords.

Assume there are question Q and candidate answer sentence W. Then, the similarity of identical keyword number is defined as follows.

$$kwSim(Q,W) = \frac{2 \times Same(Q,W)}{Num(Q) + Num(W)} \qquad (3\text{-}2)$$

where $Same(Q,W)$ represents the number of identical keywords in Q and W, and $Num(X)$ represents the number of keywords in sentence X.

The distance similarity between keywords is defined as follows.

$$disSim(Q,W) = 1 - \frac{\mathrm{Dis}(Q)}{Dis(Q) + Dis(W)} \qquad (3\text{-}3)$$

where $\mathrm{Dis}(Q)$ represents the distance between the last and the first keywords, and $Dis(W)$ the distance between the last and the first keywords of W, which are the same as those in Q.

The order similarity between keywords is defined as follows.

$$ordSim(Q,W) = 1 - \frac{\mathrm{Rev}(Q,W)}{Max\,\mathrm{Rev}(Q,W)} \qquad (3\text{-}4)$$

where $\mathrm{Rev}(Q,W)$ represents the reverse number of a natural number sequence composed of the positions of the keywords in question Q in candidate answer sentence W, and $Max\,\mathrm{Rev}(Q,W)$ represents the maximum of the above reverse number.

From the three-similarity measurement defined above, further sentence similarity based on keywords of Q and W can be obtained as:

$$\mathrm{key}Sim(Q,W) = \lambda_1 \times kwSim(Q,W) + \lambda_2 \times disSim(Q,W) + \lambda_3 \times ordSim(Q,W) \qquad (3\text{-}5)$$

where λ_1, λ_2 and λ_3 are 0.8, 0.1, and 0.1, respectively.

3.1.2. Sentence similarity method based on semantics

The semantics of words is embodied the similarity between words. The calculation of the similarity between words was based on "Synonym forests (Extended Edition)". A 5-layer hierarchical tree was formed according to the number of the words in "Synonym forests (Extended Edition)", and the similarity $WordsSim(W1,W2)$ between words $W1$ and $W2$ is calculated as follows.

When two words are not in the same tree, $WordsSim(W1,W2) = 0.1$; when they are in the same tree, $WordsSim(W1,W2) = f_i * \cos(n * \pi / 180) * [(n-k+1)/n]$,

where

- f_i represent the weights in the i-th layer, and take the values of 0.65, 0.8, 0.9 and 0.96, respectively;
- n represents the number of branches of the located layer;
- k is the distance between two branches.

A word can has multiple numbers, and the maximum is used for similarity calculation.

Let two sentences be Q and W, respectively, with the words in Q and W being $q_1, q_2, \ldots, q_m$ and $w_1, w_2, \ldots, w_n$. The word similarity between q_i and w_j is $WordsSim(q_i, w_j)$, and the semantics based sentence similarity between Q and W is given by:

$$\mathrm{ws}Sim(Q,W) = \left(\frac{\sum_{i=1}^{m} q_i}{m} + \frac{\sum_{j=1}^{n} w_j}{n} \right) / 2 \qquad (3\text{-}6)$$

where

$$q_i = \max\left(WordsSim\left(q_i, w_1\right), WordsSim\left(q_i, w_2\right), \ldots, WordsSim\left(q_i, w_n\right)\right)$$
$$w_j = \max\left(WordsSim\left(q_1, w_j\right), WordsSim\left(q_2, w_j\right), \ldots, WordsSim\left(q_m, w_j\right)\right)$$

3.1.3. BoW based sentence similarity method

The BoW features of question and candidate answers were extracted and represented by vector space model (VSM), respectively. In VSM, the questions and candidate answers were firstly represented as vectors. For example, the questions were represented as $Q\left(q_1, q_2, \ldots, q_n\right)$, while candidate answers were represented as $W\left(w_1, w_2, \ldots, w_n\right)$. Generally, the similarity between Q and W is measured by the cosine of the angle between them. The BoW based sentence similarity between Q and W is defined as follows.

$$vsmSim(Q,W) = \frac{\sum_i^n q_i \times w_i}{\sqrt{\sum_i^n q_i^{\ 2} \times \sum_i^n w_i^{\ 2}}} \qquad (3\text{-}7)$$

3.2. Algorithm description

The answer extraction algorithm based on multi-strategy fusion is described as follows.

Algorithm 2 Multi-strategy fusion based answer extraction

Input: Question set $Q = \left\{Q_i \mid i = 1, 2, \ldots n\right\}$ and the candidate answer set of each question Q_i, $W_i = \left\{W_{ij} \mid j = 1, 2, \ldots, k\right\}$

Output: The maximum sentence similarity between question Q_i and candidate answer set $W_i = \left\{W_{ij} \mid j = 1, 2, \ldots, k\right\}$

Begin

//Preprocessing stage

Extract the keywords, semantics and BoW features of each question Q_i and candidate answer sentence W_{ij}, forming the keyword set, word coding set, and BoW vector of Q_i and W_{ij}.

//Determine coefficient weights α_1、α_2、α_3

Set the initial numbers of correct answers extracted by keywords, semantics, BoW as $N_1 = 0$, $N_2 = 0$, $N_3 = 0$

For i=1 to n

For j=1 to k

Respectively calculate the similarity between Q_i and W_{ij} based on keywords, semantics, BoW, obtaining $keySim(Q_i, W_{ij})$, $wsSim\left(Q_i, W_{ij}\right)$ and $vsmSim\left(Q_i, W_{ij}\right)$

End if

If sentence with the maximum similarity based on keywords in question Q_i and candidate answer set $W_i = \left\{W_{ij} \mid j = 1, 2, \ldots, k\right\}$ is correct answer, then $N_1 ++$; likewise, whether N_2 and N_3 are incremented by 1 can be determined.

End for

Result in the accuracy of answer extraction algorithm based on keywords, semantics, BoW, $P_1 = N_1 / n$, $P_2 = N_2 / n$ $P_3 = N_3 / n$, respectively; then result in $\alpha_1 = P_1 / \left(P_1 + P_2 + P_3\right)$, $\alpha_2 = P_2 / \left(P_1 + P_2 + P_3\right)$, $\alpha_3 = P_3 / \left(P_1 + P_2 + P_3\right)$

//Use multi-strategy fusion based answer extraction algorithm to calculate sentence similarity

For any question Q_i and its candidate answer set $W_i = \left\{W_{ij} \mid j = 1, 2, \ldots, k\right\}$

For j=1 to k

Respectively calculate the sentence similarity between Q_i and W_{ij} based on keywords, semantics, BoW, resulting in $keySim(Q_i, W_{ij})$, $wsSim\left(Q_i, W_{ij}\right)$ and $vsmSim\left(Q_i, W_{ij}\right)$

$$allSim\left(Q_i, W_{ij}\right) = \alpha_1 \times keySim(Q_i, W_{ij}) + \alpha_2 \times wsSim\left(Q_i, W_{ij}\right) + \alpha_3 \times vsmSim\left(Q_i, W_{ij}\right)$$

End if

Return the maximum sentence similarity MAX_{allSim} between question Q_i and candidate answer set $W_i = \left\{W_{ij} \mid j = 1, 2, \ldots, k\right\}$

End

4. EXPERIMENT

4.1. Question classification

4.1.1. Experimental data

In our experiments, enterprise law field question dataset ELQ_set was used, in which each question had been manually labeled in establishment stage. The dataset was divided into training set and test set with a ratio of 3:1, the training set containing 654 samples, and the test set containing 219. After using word segmentation system ICTCLAS designed by Chinese Academy of Sciences for word segmentation and part of speech tagging, the basic features were extracted, including BoW, part of speech (PoS) and word semantics (WS). The features were combined (BOW/POS, BOW/WS and BOW/POS/WS), and all kinds of features were converted into the formats required by naive Bayesian (NB) algorithm and SVM.

4.1.2. Evaluation criteria

The commonly used evaluation criterion of question classification is accurate ratio (P), with the following definition

$$P = \frac{c}{n} \times 100\% \qquad (4\text{-}1)$$

where n represents the total number of question categories, and c represents the total number of correctly classified questions.

4.1.3. Experimental results and analysis

To verify the effectiveness of SVM algorithm in enterprise law field question classification, comparative experiments were conducted for it and NB, with the results of question classification experiments listed in Table 1.

Table 1: Results of question classification experiments (Accuracy/%)

Category	Human		Number		Description		Enumeration		Definition		Total Accuracy	
Method	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB
BOW	81.8	68.2	79.2	70.8	80.0	68.9	83.1	72.3	81.0	69.8	81.3	70.3
WS	68.2	63.6	70.8	66.7	71.1	66.7	70.8	67.7	68.3	66.7	69.9	66.7
POS	18.2	13.6	16.7	16.7	20.0	20.0	21.5	18.5	20.6	17.5	20.1	17.8
BOW/POS	77.3	68.2	83.3	70.8	82.2	73.3	84.6	73.8	84.1	71.4	83.1	72.1
BOW/WS	72.7	63.6	75.0	66.7	75.6	66.7	78.5	70.8	77.8	69.8	76.7	68.5
BOW/POS/WS	81.8	68.2	87.5	75.0	84.4	73.3	84.6	75.4	82.5	73.0	84.0	73.5

The experimental results in Table 1 show that the classification accuracy of using combined features is higher than that of using single features, with the highest classification accuracy obtained by BOW/POS/WS, indicating that combining the three basic features can improve the classification accuracy. Meanwhile, SVM based classification accuracy is 10 percentage points higher than the accuracy of NB, indicating that SVM based classification algorithm is effective in enterprise law question classification.

4.2. Answer extraction

4.2.1. Experimental data

To determine the parameters α_1, α_2 and α_3 in (3-1), the three basic answer extraction algorithms should be firstly tested. A total of 173 questions were selected from the question set ELQ_set and used as a test set, including 45 in description type, 65 in enumeration type, and 63 in definition type. Meanwhile, by analyzing law text, it was known that laws are in the form of clauses, and the answers to the questions generally appear in the clauses of laws. Therefore, each law clause can be treated as a paragraph, and all paragraphs form the candidate answer set $\mathbf{W_i}$. Thus, the test set with 173 questions and the candidate answer set $\mathbf{W_i}$ form the test data of this experiment.

After preprocessing including word segmentation, part of speech tagging, feature extraction, keyword extraction, etc., the formats required by all algorithms were obtained. In keyword based answer extraction, it was required that the data format be the keyword set of each question and candidate answer sentences. The data format of semantics based answer extraction was the word number (synonym forests number) set of each question and candidate answer sentences. The data format of BoW answer extraction was the BoW vector of each question and candidate answer sentences.

4.2.2. Evaluation criteria

The commonly used evaluation criteria of answer extraction are MRR criterion and answer extraction correction rate. The MRR criterion is defined as follows.

$$MRR = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} score(i) \qquad (4\text{-}2)$$

where n is the number of total test question, and the definition of $score(i)$ is as follows?

$$score(i) = \begin{cases} \dfrac{1}{CorrectRank} & \text{, There is a correct answer in the candidate answer} \\ 0 & \text{, There is no correct answer in the candidate answer} \end{cases} \qquad (4\text{-}3)$$

where $CorrectRank$ represents the position of the correct answer in the ordered candidate answer list.

The answer extraction correct ratio is defined as follows.

$$P = \frac{\text{The number of questions that correctly return the answer}}{\text{Total number of questions}} \times 100\% \qquad (4\text{-}4)$$

4.2.3. Experimental results and analysis

To accurately calculate MRR and P, in conducting the experiments, the first 5 candidate answers were considered according to the similarity level. The experimental results of using answer extraction algorithm based on the sentence similarity of KW, WS, and BoW are listed in Table 2.

Table 2: Results of three basic answer extraction algorithms

Question Category	No. of questions	MRR			P (%)		
Method		KW	WS	BOW	KW	WS	BOW
Description	45	0.543	0.716	0.636	45.0	65.0	51.7
Enumeration	65	0.597	0.731	0.671	52.5	67.5	57.5
Definition	63	0.565	0.728	0.648	49.2	65.8	54.7
Total	173	0.580	0.727	0.651	49.7	66.3	55.0

It is clear from the experimental results in Table 4-2 that the accuracy of semantics based answer extraction is the highest, and the three algorithms are effective for answer extraction. The total accuracy of the three algorithms was used as the basis for determining parameters α_1, α_2 and α_3 in (3-1), with the calculation process as follows.

$$\alpha_1 = \frac{49.7}{49.7+66.3+55.0} = 0.29$$
$$\alpha_2 = \frac{66.3}{49.7+66.3+55.0} = 0.39$$
$$\alpha_3 = \frac{55.0}{49.7+66.3+55.0} = 0.32$$

Substituting parameters α_1, α_2 and α_3 in (3-1), and (3-1) is reformulated as:

$$allSim(Q,W) = 0.29 \times keySim(Q,W) + 0.39 \times wsSim(Q,W) + 0.32 \times vsmSim(Q,W)$$

Experiments was carried out for the proposed multi-strategy fusion based answer extraction algorithm, and the experimental results shown in Table 3 were obtained.

Table 3: Experimental results of multi-strategy fusion based answer extraction algorithm

Question Category	Number	MRR	P (%)
Description	45	0.785	68.3
Enumeration	65	0.781	75.8
Definition	63	0.776	72.5
Total	173	0.791	73.0

It is clear in the experimental results that the MRR of multi-strategy fusion based answer extraction algorithm is 0.791, and the accuracy rate is 73%. Compared with other algorithm, the performance is largely improved, indicating the effectiveness of this algorithm in enterprise law field answer extraction.

5. CONCLUSION

In this study, for question classification in enterprise law field, comparative experiments were conducted for SVM based classification method and NB based classification method. Experimental results show that SVM based classification method has higher accuracy. For enterprise law field answer extraction, the multi-strategy fusion based answer extraction algorithm was proposed in this study, which combined keyword, semantics, and BoW answer extraction strategies. Experiments were carried out to verify the effectiveness of the algorithm. Experimental results demonstrate that the proposed methods can be well applied in the intelligent question answering system in enterprise law field.

REFERENCE

[1] Marc, L., Gideon, S.M., Ellen, R., Breck, E. 2001. Analyses for elucidating current question answering technology. Journal of Natural Language Engineering, 7 (4), 325-342.

[2] Dwivedi, S.K., Singh, V. 2014. Integrated question classification based on rules and pattern matching [C]. Proceedings of the 2014 international conference on information and communication technology for competitive strategies, 1-7.

[3] Yue, C.Y., Cao, H.Q., Xiong, K., Cui, A., Qin, H., Li, M. 2017. Enhanced question understanding with dynamic memory networks for textual question answering [J]. Expert systems with applications, (80), 39-45.

[4] Lee, W.S., Zhang. D. 2003. Question classification using support vector machines[C]. Proceedings of the international ACM SIGIR conference on research and development in information retrieval, 26-32.

[5] Li, X., Roth, D. 2002. Learning question classifiers [J]. Coling, 12 (24), 556-562.

[6] Derici, C., Celik, K., Ozgur, A., Gungor, T., Kutbay, E., Aydin, Y., Kartal, G. 2014. Rule-Based Focus Extraction in Turkish Question Answering Systems [C]. Signal Processing and communications applications conference, 1604-1607.

[7] El-Ansari, A., Beni-Hssane, A., Saadi, M. 2017. A multiple ontology based system for answering natural language questions [C]. Europe and MENA cooperation advances in information and communication technologies, (520), 177-186.

[8] Ramprasath, M., Hariharan, S. 2016. Algorithm for answer extraction based on pattern learning [J]. The international arab journal of information technology, 13 (1ASI), 178-183.

[9] Baazaoui-Zghal, H., Besbes, G. 2014. Ontology-based answer extraction method [J]. Lecture Notes in Computer Science, (8741), 239-240.

[10] Jayakodi, K., Bandara, M., Meedeniya, D. 2016. An automatic classifier for exam questions with WordNet and Cosine similarity [C]. Moratuwa

Engineering Research Conference, 12-17.

[11] Bang, S.J., Wu, W. 2016. Naive Bayes Ensemble: A new approach to classifying unlabeled multi-class asthma subjects [C]. International conference on bioinformatics and biomedicine, 460-465.