Indian Journal of Science and Technology, Vol 9(17), DOI: 10.17485/ijst/2016/v9i17/93160, May 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Improving Question Classification by Feature Extraction and Selection

Nguyen Van-Tu[1], Le Anh-Cuong[2*]

[1]VNU University of Engineering and Technology, Ha Noi City, Vietnam; tuspttb@gmail.com
[2]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam; leanhcuong@tdt.edu.vn

Abstract

Question classification is the task of predicting the entity type of the answering sentence for a given question in natural language. It plays an important role in finding or constructing accurate answers and therefore helps to improve quality of automated question answering systems. Different lexical, syntactical and semantic features was extracted automatically from a question to serve the classification in previous studies. However, combining all those features doesn't always give the best results for all types of questions. Different from previous studies, this paper focuses on the problem of how to extract and select efficient features adapting to each different types of question. We first propose a method of using a feature selection algorithm to determine appropriate features corresponding to different question types. Secondly, we design a new type of features, which is based on question patterns. We tested our proposed approach on the benchmark dataset TREC and using Support Vector Machines (SVM) for the classification algorithm. The experiment shows obtained results with the accuracies of 95.2% and 91.6% for coarse grain and fine grain data sets respectively, which are much better in comparison with the previous studies.

Keywords: Feature Extraction, Feature Selection, Question Answering Systems, Question Classification, Question Patterns

1. Introduction

Automated Question Answering has become an important research direction in natural language processing[1,2]. Its purpose is to seek an accurate and concise answer to a free-form factual question from a large collection of text data, rather than a full document, judged relevant as in standard information retrieval tasks. Although different types of question answering systems have different architectures, most of them follow a framework in which question classification plays an important role[3]. Furthermore, some studies have demonstrated that performance of question classification has significant influence on the overall performance of a question answering system[2,4,5]. The task of question classification is to predict the entity type of the answer of a natural language question[6]. For example, for the question "Where is the Eiffel Tower?", the task of question classification is to return label "location", thus the answer to this question is a named entity of type "location". Since we predict the type of the answer, question classification is also referred as answer type prediction.

Many studies have addressed this problem, they belongs to the rule-based approach[7,8] or machine learning-based approach[4,9-11]. In this paper, we follow the machine learning approach and pay attension on the importance of feature extraction and selection. From the view of machine learning, we can easily formulate the this task as a classification problem. There are various supervised learning methods used such as Nearest Neighbors (NN), Naive Bayes (NB), Decision Tree (DT), Sparse Network of Winnows (SNoW), and Support Vector Machines (SVM). However, as expressed from experimental results in previous studies, feature sets affect much the quality of question classification.

*Author for correspondence

According to previous studies, various types of features have been investigated. The most common types are bag of words and n-grams which were used in all studies. Some other studies (e.g[12]) tried to enrich the feature set by adding more linguistic information as part-of-speech tags or head words, or even semantic features. However, from our observation combining all features is not always the best solution for all questions. Therefore, in this paper we will give an experimental investigation for finding the best feature sets corresponding to different groups of questions. In addition, we also extract a new type of features based on question patterns. These new features are then integrated to the existed feature sets and receive better results of classification. We tested our proposed feature sets using a SVM classifier which is experimental shown to get best results in[6,9,13,14]. And like most previous studies, the TREC dataset is chosen for conducting experiments.

The rest of this paper is organized as follows: Section 2 presents the basic issues in question classification including question type taxonomy and feature extraction. Section 3 presents our proposal for feature selection. Section 4 presents the experiments. Conclusion and future works will be presented in section 5.

2. Basic Issues of Question Classification

2.1. Question Type Taxonomy

The set of question categories (classes) are usually referred as question taxonomy. Different question taxonomies have been proposed in different works, but most of the recent studies used the two layer taxonomy proposed by [15][1]. This taxonomy consists of 6 coarse grained classes and 50 fine grained classes. Table 1 lists this taxonomy.

Whenever the entity of answering is determined we can combine it with other information to find correct answers. For example, if we know the question is asking about location (or more concrete, a city), it is easier to find the exact information for answering as well as to form the appropriate answer.

2.2. Classification Algorithms and Evaluation

Machine Learning Approach

Most studies in question classification follow supervised machine learning approach. There are many different classification methods used such as: Support Vector Machine, Naive Bayesian classification, Maximum Entropy Models[16,17], Sparse Network of Winnows[12]. Among these methods, Support Vector Machine with linear kernel function is shown as the most effective method, according to[6,9,13,14]. Therefore, SVM is the machine learning method used in our system. We can easily search for a many documents introducing about SVM methods and applications, thus it is not necessary for presenting it in detail here.

A general framework in supervised machine learning method for question classification is briefly described in the following steps:

- First, we need to build a training dataset, it includes questions assigned with classification labels.
- Second, each labeled question in the training dataset is represented as a vector of features.
- Third, a machine learning method (here SVM) is used to learn on the training vectors and generate the classifier.
- Finally, for each a test question we represent it by a vector of features and use the learnt classifier to obtain a label (i.e. a question category).

Table 1. The coarse and fine grained question classes

Coarse	Fine
ABBREVIATION	Abbreviation, expression
ENTITY	Animal, body, color, creative, currency, dis.med, event, food, instrument, lang, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
DESCRIPTION	Definition, description, manner, reason
HUMAN	Group, individual, title, description
LOCATION	City, country, mountain, other, state
NUMERIC	Code, count, date, distance, money, order, other, period, percent, speed, temperature, size, weight

Classification Evaluation

Performance in question classification is evaluated by the global accuracy of the classifier for all the coarse or fine classes[12].

$$\text{Accuracy} = \frac{\#\text{of correct predictions}}{\#\text{of predictions}}$$

There is also the accuracy of a question classifier on a specific class precision. Precision in question classification on a specific class c is defined as follows:

$$\text{Precision(c)} = \frac{\#\text{of correct predictions of class c}}{\#\text{of predictions of class c}}$$

For the systems in which a question can only have one class, a question is correctly classified if the predicted label is the same as the true label. But for the systems which allow a question to be classified in more than one class [12,15], a question is correctly classified, if one of the predicted labels is the same as the true label.

2.3. Feature Extraction

There are various types of features which are currently used for question classification. They can be grouped into three different categories based on the kinds of linguistic information: lexical, syntactical and semantic features.

2.3.1 Lexical Features

Lexical features are usually the context words appearing in the question. In question classification, a question is represented similarly to document representation in the vector space model, i.e., a question can be represented as the vector:

$$q = (q_1, q_2, \ldots, q_N)$$

where q_i is defined as the frequency of term i^{th} in question q and N is the total number of terms. Note that only non-zero valued features are kept in the feature vector. Then, a question q is represented in the form:

$$q = \{(t_1, f_1), \ldots, (t_p, f_p)\}$$

where f_i is the frequency of the term i^{th} in the question q. These features are called bag-of-words features or unigrams features.

Unigrams is a special case of the so-called n-gram. To extract n-gram features, any n consecutive words in a question is considered as a feature. Table 2 lists the lexical features of the sample question "Who was elected president of South Africa in 1994?"

Note that some special cases of getting lexical information like question words (i.e: who, how, when, what) or word-shapes are put into the lexical feature set.

2.3.2. Syntactic Features

Syntactic features are extracted from the syntactical structure of a question. There are two common kinds of syntactic features used for question classification, including *tagged unigrams* and *head words*.

Tagged Unigrams

Tagged Unigrams indicate the part-of-speech tag of each word in a question like NN (Noun), NP (Noun Phrase), VP (Verb Phrase), JJ (adjective), and etc. The following example shows these features extracted from the sentence "Who was elected president of South Africa in 1994?"

{Who_WP, was_VBD, elected_VBN, president_NN, of_IN, South_NNP, Africa_NNP, in_IN, 1994_CD,?.}

Head Words

A head word is considered as the key word or the central word in a sentence, a clause or a phrase. This word is determined based on the syntactic parsed tree of the input sentence. As mentioned in[6], Head Words contain important information for specifying the object that a question is seeking. Therefore, identifying the head word correctly can improve the classification accuracy since it is the most informative word in the question.

Table 2. Example of lexical features

Feature Space	Features
Unigram	{(Who, 1) (was, 1) (elected, 1) (president, 1) (of, 1) (South, 1) (Africa, 1) (in, 1) (1994, 1) (?, 1)}
Bigram	{(Who-was, 1), (was-elected, 1), (elected-president, 1), (president-of, 1), (of-South, 1), (South-Africa, 1), (Africa-in, 1), (in-1994, 1), (1994-?, 1)}
Trigram	{(Who-was-elected, 1), (was-elected-president, 1), …, (in-1994-?, 1)}
Wh-Word	{(Who, 1)}
Word-Shapes	{(lowercase, 5) (mix, 3) (digit, 1) (other, 1)}

Table 3. A sample of questions with their headwords and appropriate categories

Question	Category
What city has the zip code of 35824 ?	LOC:city
Who developed the vaccination against polio ?	HUM:ind
Who invented the slinky ?	HUM:ind
George Bush purchased a small interest in which baseball team ?	HUM:gr
When did Idaho become a state ?	NUM:date
What river flows between Fargo, North Dakota and Moorhead, Minnesota ?	LOC:other
What is the oldest city in Spain ?	LOC:city

For example, for the question "What is the oldest city in Spain?" the head word here is "city". The word "city" in this question can highly contribute to the classifier for classifying this question as "LOC:city". Table 3 lists sample questions from TREC dataset together with their class labels. The head words are identified by being underlined.

To determine the head word of a sentence, a syntactic parser is required. For sentences written in English language, people usually use the Stanford PCFG parser[18] which is also used in this paper.

2.3.3. Semantic Features

Semantic features are useful in the case of sparse data. From higher level semantic concept we can get the relationship (i.e. the semantic similarity) between different words. WordNet is a well-known resources used for determining semantic features. WordNet is a lexical database of English words providing a lexical hierarchy that associates a word with higher level semantic concepts namely hypernyms[6]. For example a hypernym of the word "city" is "municipality".

There are three kinds of semantic features being used for question classification, as shown in[12], as follows:

Question Category (QC)

WordNet hierarchy is used to estimate the similarity of question's head word. The class with highest similarity is considered as a new feature and will be added to the feature vector. For example, the question "What American composer wrote the music for "West Side Story"?" has its head word "composer". To find the question category feature, the similarity of this word will be compared with the similarity of all question categories. The category with the highest similarity will be added to the feature vector. In this example the most similar category is "individual" and therefore the question category feature will be {(individual, 1)}.

Question Expansion (QE)

Another semantic feature called query expansion which is basically very similar to hypernym features. As we explained before, we add hypernym of a head word to the feature vector with words from WordNet hierarchy. Instead of imposing this limitation, we defined a weight parameter which decreases by increasing the distance of a hypernym from the original word. For example for the question "What river flows between Fargo, North Dakota and Moorhead, Minnesota?". The head word of this question is "river". The query expansion features of this question will be as follows, given that the weight of "river" is considered as 1:

{(river, 1) (stream, 0.6) (body-of-water, 0.36) (thing, 0.22) (physical-entity, 0.13) (entity, 0.08)}.

Related Words (RW)

Another semantic feature that we also use in this work is the related words as presented in[12]. In this study, the authors defined groups of words, each was represented by a category name. If a word in the question exists in one or more groups, its corresponding categories will be added to the feature vector. For example if any of the words {birthday, birthdate, day, decade, hour, week, month, year} exists in a question, then its category name, "date", will be added to the feature vector.

Table 4 lists semantic features the question "What river flows between Fargo, North Dakota and Moorhead, Minnesota?".

Table 4. Example of semantic features

Feature Space	Features
Hypernyms	{(river, 1) (stream, 1) (body-of-water, 1) (thing, 1) (physical-entity, 1) (entity, 1)}
Related Words	{(rel:What, 1) (rel:list.tar, 2) (rel:loca, 2)}
Question Category	{(other, 1)}
Query Expansion	{(river, 1) (stream, 0.6) (body-of-water, 0.36) (thing, 0.22) (physical-entity, 0.13) (entity, 0.08)}

3. Our Proposal of Feature Selection and Adding more New Feature Type

3.1 Combination of Different Feature Sets

Suppose that each feature type as mentioned in section 2.3 will generate a single set of features. A natural way for obtaining the final set of features to use in the question classification is to combine all the single sets. However, we found that the combination of all these feature sets is not efficient and doesn't always give the best results for all questions.

From our observation, we can recognize that each type of questions can be sensitive with particular types of features. Therefore, assigning different feature sets corresponding with different question types can be a solution. The question types here relate to the question words: "who", "when", "how", "why", "which", "where", and "what". It means each the question word defines one feature type. We reserve one type for the remaining questions which do not contain those question words.

We propose to use a simple feature selection for determining the best combination of single feature sets for each question types, as presented in the algorithm 1 below:

Algorithm 1. Determining the feature set for each question type

Input: a training data and a development data set corresponding to the selected feature type; a learning machine method (e.g SVM here)

Output: a set of single feature types which gives the best result on the development data set.

Step 1: extract all the single feature sets, denoted by SF_1, SF_2, …, SF_n set the remain feature sets SF = { SF_1, SF_2, …, SF_n }
set the initial feature set F = { }
set the intitial accuracy $A = 0$

Step 2: For each SF*i* in SF train a new classifier again with the new feature set F+{SF_i} ; get the accuracy tested on the development test, denote it by A_i

Step 3: get A_k to be the highest accuracy, corresponding to the feature set F+{SF_k};
If $A_k > A$
set $A = A_k$
SF = SF\{SF_k}
F = F + { SF_k }
Else
Return F; Quit

Step 4: If SF is not empty Repeat at Step 2
Else
Return F; Quit

3.2 Extracting Features from Question Patterns

By studying TREC dataset we found some questions inherently do not have any head word. For example, the sentence "What is an atom?" has no suitable head word as the entity type of the only noun ("atom") in this question. It does not provide necessary information to classify this question as "definition". We recognize that by integrating lexical, syntactic and semantic information into an unique form, we can get richer features for determining correct labels of such questions. This new kind of features also bring advanced evidences to the classification and therefore may lead to a better result.

We first design some patterns (i.e. templates) for containing the integrated of lexical, syntactic, and semantic information. Table 5 shows some designed question patterns.

From these patterns which we call question patterns, we will generate corresponding features. For example, from the question "How is thalassemia defined?" we can be received the features (How-is, 1) and (How-is-defined, 1). We then combine these features with the existed feature sets to get the final feature sets for classification.

Table 5. Example of question patterns

Question patterns	Explain the semantic information
Wh-word + Tobe + word-shape	
Wh-word + weather-word	weather word: hot, cold, warm, wet, …
Wh-word + distance-word	distance-word: far, long, …
Wh-word + Tobe + distance-word	distance-word: far, long, …
Wh-word + money-word	money-word: money, cost, rent, sell, spend, charge, pay, …
Wh-word + place-word	place-word: city, county, mountain, state,…
Wh-word + reason-word	reason-word: causes, used, known, …

4. Experiments and Results

The dataset we used for conducting our experiment was created by[15]. They provided a question dataset which is widely used in question classification studies and known as UIUC or TREC[2] dataset. It consists of 5500 labeled questions which is used as training set and 500 indepen-

dent labeled questions which is used as the test set. The 5500 training questions are split randomly into 5 different training sets with the size 1000, 2000, 3000, 4000 and 5500 respectively.

We design different experiments as follows.

A. Experiment 1

For the first experiment we combine all the single feature sets for the task of classification, that includes: Unigram (U), Bigram (B), Wh-Word (WH), Word-Shapes (WS), Head-Word (H), Query-Expansion (QE), Question-Category (QC), Related-Words (R). Table 6 shows the results corresponding with different training data sets.

Table 6. The accuracy of using SVM classifier with combining the feature kinds: U, B, WH, WS, H, QE, QC, R

Training size	1000	2000	3000	4000	5500
Coarse 1	90.20%	91.20%	92.00%	92.60%	94.20%
Fine 1	79.00%	85.40%	86.60%	88.00%	90.40%

Experiment 2

In this experiment we would like to examine the contribution of the question pattern features by adding the these QP features to the feature set from the Experiment 1. Its results are shown in the Table 7.

Table 7. Accuracy of using more Question-Pattern feature

Training size	1000	2000	3000	4000	5500
Coarse 2	90.40%	91.40%	92.80%	93.20%	95.00%
Fine 2	79.20%	86.00%	87.00%	88.60%	91.00%

Comparing results of experiment 1 and experiment 2, we can see that the QP feature set actually improves the accuracy of question classification for all the training data sets.

C. Experiment 3

This experiment implements the Algorithm 1 for feature selection. Note that the QP feature set is also considered as one single feature set which is used in the algorithm. Table 8 also shows the selected feature types for each kind of question. It is worth to emphasize that the QP feature set is selected for the questions containing Wh-words, but not for the other type of questions. It seems reasonable because the QP features are designed to contain Wh-words, therefore they don't affect the question without Wh-words.

Table 8. Result of feature selection

Question types	Features
How, Who, Why, When, Where, Which	Unigram, Bigram, Word-Shapes, Question Pattern
What	Unigram, Bigram, Head word, Word-Shape, Related Words, Question Pattern, Query Expansion, Question Category
Other questions	Unigram, Bigram, Word-Shape, Related Words

Table 9 shows the accuracy of question classification when using result of the feature selection.

Table 9. Accuracy of using feature selection corresponding to question types

Training size	1000	2000	3000	4000	5500
Coarse 3	90.40%	91.60%	93.20%	93.80%	95.20%
Fine 3	79.20%	86.60%	87.40%	89.00%	91.60%

Figure 1 and figure 2 show the comparisons of the experiment 2 and experiment 3 for the coarse and fine grained question classes.

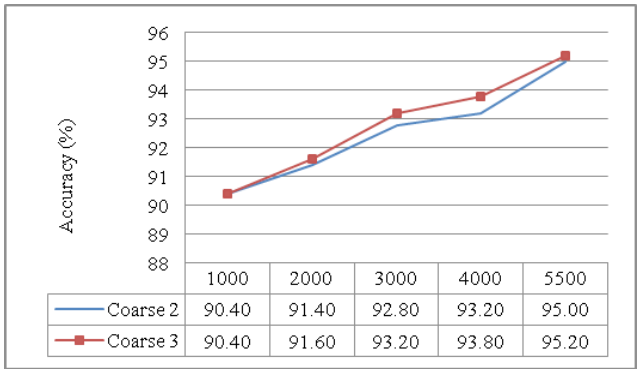

Figure 1 Result for Coarse classes in the experiment 2 and the experiment 3.

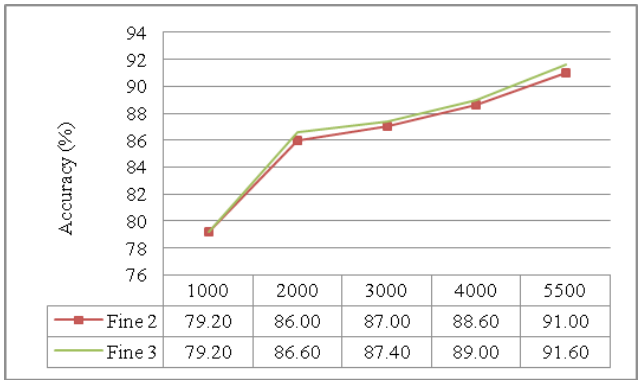

Figure 2 Result for fine grained classes in the experiment 2 and the experiment 3.

Table 10. Comparison with previous studies for the same task and the same data set

Study	Classifier	Features	Accuracy	
			Coarse	Fine
Li and Roth (2002) [15]	SNoW	U+P+HC+NE+R	91.0%	84.2%
Li and Roth (2004) [12]	SNoW	U+P+HC+NE+R+S		89.3%
Metzler et al. (2005) [10]	RBF kernel SVM	U+B+H+HY	90.2%	83.6%
Huang et al. (2008) [6]	Linear SVM	U+WH+WS+H+HY+IH	93.4%	89.2%
Silva et al. (2011) [11]	Linear SVM	U+H+C	95.0%	90.8%
Hardy et al. (2013) [19]	Extreme Learning Machine	WH + H + HY	92.8%	84.6%
Our work	Linear SVM	U+B+WS+H+R+QE+QC+QP	95.2%	91.6%

Comparing result from Table 9 with results from Table 6 and Table 7, and as illustrated in Figure 1 and Figure 2 we can see that combining both solutions (using QP features and using feature selection) significantly improves the task of question classification.

Comparison with previous studies:
In addition, we also make a comparison with well-known previous studies of this task which also used the same data set. The Table 10 shows the accuracy for the Coarse classes and the Fine grained classes.

Table 10 shows that our proposal achieve the accuracies of 95.2% and 91.6% for coarse grain and fine grain respectively, which are much better in comparison with the previous studies.

5. Conclusion

In this paper we have presented our proposal of feature extraction and feature selection for improving question classification. We have investigated various types of features including lexical, syntactic, and semantic features. We also proposed a new type of feature based on question pattern and then applying a feature selection algorithm to determine the most appropriate feature set for each type of questions. The experimental results shows that our proposal gives the best accuracies for both the Coarse classes and the Fine grained classes of questions, in comparison with using the conventional feature set, as well as in comparison with the previous studies.

6. Acknowledgments

This work is supported by the Nafosted project 102.01-2014.22

7. References

1. Wendy G Lehnert. A conceptual theory of question answering. In Proceedings of the 5th international joint conference on Artificial intelligence. 1977; 1:158–64.
2. Moldovan Dan, Pasca Marius, Harabagiu Sanda and Surdeanu Mihai. Performance issues and error analysis in an open-domain question answering system. ACM Trans, Inf. Syst. 2003; p. 133–54.
3. Ellen M. Voorhees. Overview of the trec 2001 question answering track. In Proceedings of the Tenth Text REtrieval Conference (TREC). 2001; p. 42–51.
4. Hermjakob Ulf, Hovy Eduard and Lin Chin Yew. Automated question answering in webclopedia - a demonstration. In Proceedings of ACL-02. 2002.
5. Ittycheriah A, Franz M, Zhu WJ, Ratnaparkhi A and Mammone RJ. IBM's statistical question answering system. NIST, In Proceedings of the 9th Text Retrieval Conference. 2001.
6. Zhiheng Huang, Thint Marcus and Zengchang Qin. Question classification using head words and their hypernyms. EMNLP '08, In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008; p. 927–36.
7. David A Hull. Xerox TREC-8 question answering track report. In Voorhees and Harman. 1999.
8. Prager John, Radev Dragomir, Brown Eric and Coden Anni. The use of predictive annotation for question answering in trec8. In NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8). 1999; p. 399–411.
9. Zhiheng Huang, Thint Marcus and Celikyilmaz Asli. Investigation of question classifier in question answering. EMNLP '09, In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009; p. 543–50.
10. Metzler Donald and Bruce Croft W. Analysis of statistical question classification for fact-based questions. Inf. Retr. 2005; p. 481–504.

11. Silva Joao, Coheur Luısa, Mendes Ana and Wichert Andreas. From symbolic to subsymbolic information in question classification. Artificial Intelligence Review. 2011; p. 137–54.
12. Li Xin and Roth Dan. Learning question classifiers: The role of semantic information, In Proc, International Conference on Computational Linguistics (COLING). 2004; p. 556–62.
13. Krishnan Vijay, Das Sujatha and Chakrabarti Soumen. Enhanced answer type inference from questions using sequential models. HLT '05, In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005; p. 315–22.
14. Zhang Dell and Lee Wee Sun. Question classification using support vector machines. SIGIR '03, In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 2003; p. 26–32.
15. Li Xin and Roth Dan. Learning question classifiers. COLING '02, In Proceedings of the 19th international conference on Computational linguistics. 2002; p. 1–7.
16. Blunsom Phil, Kocik Krystle and James R. Curran. Question classification with loglinear odels. SIGIR '06, In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006; p. 615–16.
17. Xin Li, Xuan-Jing Huang and Li-De Wu. Question classification using multiple classifiers. In Proceedings of the 5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network. 2005.
18. Petrov Slav and Klein Dan. Improved inference for unlexicalized parsing. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference. 2007; p. 404–11.
19. Hardy, Cheah Yu-N. Question Classification Using Extreme Learning Machine on Semantic Features. J. ICT Res. Appl. 2013; 7(1):36-58.