Journal of Theoretical and Applied Information Technology
15th November 2018. Vol.96. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PERSIAN QUESTION CLASSIFICATION USING HEADWORD AND SEMANTIC FEATURES

AMIR ROUSTAEI[1], HAMID RASTEGARI[2,*]

[1]Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
[2]Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
Email: [1]amir.ro62@yahoo.com, [2]rastegari@iaun.ac.ir

*Corresponding Author: Hamid Rastegari, Email: rastegari@iaun.ac.ir

ABSTRACT

Question classification is an important component in question answering systems. The task of question classifier is to assign a label, depending on the classification strategy, to written question in natural language. Features are essential elements to obtaining an accurate question classifier. Low accuracy at the fine-grained level is the main problem among classifiers. In this paper, in order to improve the accuracy of question classification, two new features such as question's headword and related semantic words are introduced. If headword is correctly identified, then the accuracy of answer classification increases. On the other hand, semantic meaning of related words effects on accuracy of the answer classification for both coarse and fine grained classes. The result shows the contribution of the presented features in coarse- and fine-grained classification accuracy.

Keywords: *Question Answering, Questions Classification, Machine Learning, Feature Extraction, Headword, Coarse and Fine-Grained Classification*.

1. INTRODUCTION

With the rapid increase in the amount of knowledge on the web, search engines must be more intelligent than ever before. In many cases, instead of a list of documents, a user needs only valuable information. Reading all documents to find the answer to the user's question is a time-consuming task. Therefore, a short and concise answer is preferable [1,2]. Hence, auto-answering systems satisfy this need. On the other hand, the first and most important component of such systems is question classification, the output of which plays a considerable role in finding correct and precise answers [3,4]. Hence, creating a high-precision question classifier is essential for the performance of the answering systems.

Question target classification (QTC) and subjective question classification are two types of question classification that differ noticeably in purpose. The former is used to classify questions in accordance with answer categories while the latter, classifies questions according to relevant topics [5]. Although research on the subjectivity of questions may help to identify the purpose of the question classification, the QTC plays a more important role in determining the answer type, which is more useful for filtering out irrelevant answer candidates. Therefore, the classification system in this study is based on QTC.

Questions classification should cover all kinds of questions. Moreover, it has to be balanced in terms of simplicity and power. In other words, if the classification includes a smaller number of classes than answers then it will be general, repetitive, and boring. On the other hand, if the number of classes is increased then the complexity of the system will also increase, causing difficulties in the processing and analysis of the answer. Accordingly, the use of hierarchical classification is recommended so that the number of first-level classes is low. Each first-level class incorporates a number of subclasses. This ensures more comprehensive classification. The first-level classes are referred to as "coarse-grained" while the sub-classes of the second level are called "fine-grained" [6].

Journal of Theoretical and Applied Information Technology
15th November 2018. Vol.96. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

There are two main approaches to question classification: rule-based and machine-learning-based approaches [7,8]. A rule-based approach tries to match questions using hand-written rules. This approach, however, suffers from the need to define a large number of rules. In addition, while the rule-based approach may perform well on some datasets, it may perform very poorly with others, and thus, its scalability is difficult [9]. While a considerable number of initial studies have used a rule-based approach for question classification, successful systems are often based on learning methods [3,10]. A machine-learning approach is used for question classification in this study. The main difficulty of most classifications is the low accuracy on a fine grained level. Question classifier with high precision in the answer systems is necessary. On the other hand, there has not been much work in Persian in this regard, and the only significant work done is to categorize questions with specific subject areas and use limited syntactic and lexical features. Therefore, due to the importance and application of the topic and the lack of work done for the Persian language, further researches in this field are needed. In this paper, a target-oriented question classifier is presented for use with a Persian language automatic Q/A system with a supervised approach. Since the question classification determines the type of answer, a more accurate answer leads to a more limited retrieval of documents based on the type of question, and eventually, the final answer is more accurately obtained. Thus, considering the importance of the output answer type in finding precise answers to questions, the aim of this study is to improve the classifier's accuracy for Q/A systems by extracting the question's headword and obtaining the relevant semantic words feature. According to the literature, the two-level classification (which includes the coarse- and fine-grained category) has been more successful, so it was selected here [8].
Question headword has very particular importance since it has been used to extract its feature and semantic words characteristics. Based on most of the articles in the literature, fine classification delivers poorer accuracy than coarse classification. Therefore, in this study, the fine classification accuracy is improved by utilizing the dataset containing the relevant semantic words associated with each word in the fine category and by considering the question's headword. In fact, the main weakness of all presented methods in fine grained category accuracy has not yet reached an ideal precision. We are going to work on a standard dataset that are not limited to a specific domain. On the other hand, instead of using a rule-based approach and providing multiple rules that increase time and complexity, we will extract new features that increase the accuracy of classification.
The body of the paper is organized as follows: in section II, the dataset and classification are described. The proposed method is presented in section III. In section IV, the tests and evaluation are reported upon. Finally, in section V, conclusions are discussed and potential future work is proposed.

2. DATASET

The first step for question classification is to define the classes of question or the types of questions. Because supervised machine learning approach has been used here, the dataset is required for training the classifier. Data is collected in two stages; the first stage consists of Persian question sentences, which includes both coarse and fine grained category label for each question that is manually labeled. In the second stage, using feature extraction functions, each question is converted to vector features. So, there would be a matrix in the final dataset which each row contains a feature vector of the question and the last two columns indicate the label of coarse and fine grained categories respectively.
In this research, the six coarse and fine classes are used that the label for these classes is in accordance with the classification of two layers provided by Li & Roth in 2002 [6]. The majority of studies in this field are based on this classification [3,12,13]. Persian questions are taken from Hamshahri newspaper queries set and Frequently Asked Questions (FAQ) available on multiple websites. Table 1 shows the coarse and fine grained question classification in each category.

Table 1: Coarse and Find Grained Classification

Coarse	Fine
Abbreviation	Abbreviation ،explanation
Description	definition, description, manner, reason
Entity	animal, body, color, creation, currency disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
Location	city, country, mountain, other, state

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Numeric	code, count, date, distance, money, order, other,percent, period, speed, temperature, size, weight
Human	description, group, individual, title

3. PROPOSED APPROACH

The main idea of the proposed approach is provided in this section. In Section A, system architecture is presented. The system architecture consists of three main parts; question preprocessing, feature extraction and classification using machine learning algorithms. The proposed approach model is shown in Figure 1.

3.1 Question Preprocessing

The sentences preprocessing steps consists of Persian texts normalization, words identification and stemming. For all preprocessing stages the Persian text preprocessing tool from Web Technology Laboratory of Mashhad[1] is used.

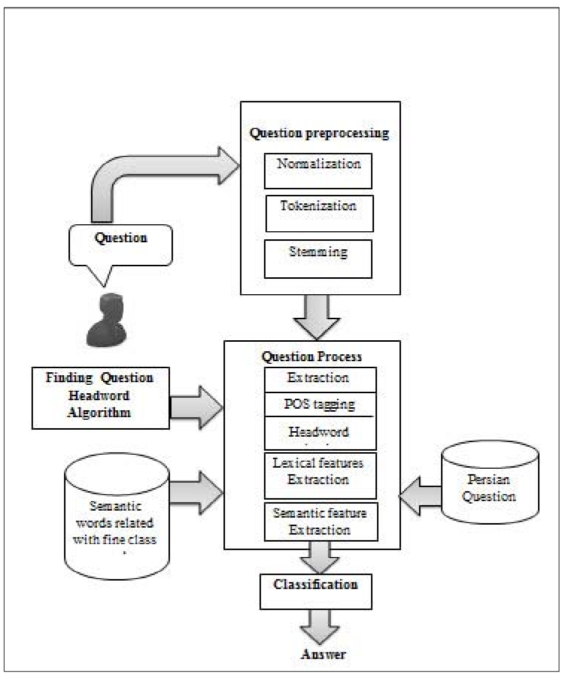

Figure 1. The Proposed Model

3.2 Question Processing

For the current task, one question sentence is exhibited as a feature vector and behaves whether as a test example or training for learning. Mapping of a question to a class label is a defined linear function on the feature vector. Features in question classification can be classified into three different types of lexical, syntactic and semantic features categories [14].

• Lexical Features

Lexical features of a question are often based on field questions words, words that appear in the question [14]. Question words characteristics, bag of words characteristics, question length and position of question word are among the lexical features which have been used in this study.
Bag of word model have been used to show a question like a collection of unordered words [9]. In this model, a dictionary of length equal to the total number of distinct words in the training set is created. Then, the question is being shown like a binary feature vector. If the question includes the i^{th} word in the dictionary, then the corresponding value is 1 otherwise it is 0.

• Syntactic Feature

Extracted features in this category are POS and headword features. The headword features specifies the target of the question and generally show the question what to follow. These words are usually extracted based on syntactic structure. So, for its extraction firstly it is needed to have the syntax label and the syntax of grammar based questions.
The idea of extracting the headword from the syntactic tree was primarily introduced by (Collins, 1999) [14]. He proposed to utilize some known rules so-called Collin rules to extract the head from the clauses. In order to find the headword of the clause, the parser tree is traversed in top-down fashion and at any level under a tree which includes the headword is identified with Collin rules. However, the Collins rules were not appropriate for question classification task, because they prefer verbal phrase on the noun phrase while for question classification cases which headword has been seeking, noun phrase should be preferred on verbal phrase. The rules for the extraction of semantic headword was amended in [15]. Other researchers redefined the rules and also they defined the regular expression patterns to help the words identification [11].

[1] https://wtlab.um.ac.ir

Huang's algorithm has better results than other works in this field and is used in most studies.
The question's headword extraction in Persian is a highly challenging task. One reason is that unfortunately parsers are very limited in Persian language and the existing parsers have not reached to acceptable accuracy level yet while in English language the variety of high-precision parsers are available. Therefore, it is not possible to rely only on grammatical rules and syntax tree traversal to extract headwords. Accordingly, we examined the question pronouns found in Persian, Persian questions and multiple placement scenarios of these pronouns in the phrase. Then, regular expressions to find the headword in question phrases in Persian are defined. The question phrase is compared with each defined regular expressions and once it was matched with the corresponding regular phrase then according to defined regular phrase the position of headword will be found and since in preprocessing stage the syntactic label of each question is identified the corresponding noun phrase can be extracted. The algorithm to find the headword is shown in figure 2.
In question phrases including question words like:
"چرا/chera/Why",
"چطور/chetor/How", or
"چگونه/chegone/How
the purpose of the question lies inside the question-word and as such the headword cannot be extracted from any word in the phrase. But these question-words are in common on how to answer the question.
The answer for these types of questions is descriptive type. Therefore, for the phrases which involve such type of question words instead of question's headword the descriptive level must be delivered. For instance, the answer type for the question like
"چگونه میتوان امنیت اطلاعات را حفظ کرد؟ /
How can preserve the information security? "
is describing a method on how to preserve the information security.
In question phrases including question-words like:
"کجا/koja/Where"
the purpose of the question is to find the name of the place. For this case also the headword cannot be seen in the question phrase. Therefore, to improve the algorithm accuracy, the corresponding level type which is the "position" will be delivered as a headword.

Pattern 1: Phrases containing question pronouns of objects, people, time or quantity, the same name after the word
"چه/che/What"
will be delivered as a question's headword.

Pattern 2: In question phrases if after the question pronoun of
"چه/che/What"
a name is appeared then, the first noun in the phrase is extracted as a headword.

Pattern 3: If after the question pronoun of
"چه/che/What"
the linking verb is appeared then, the first noun in the phrase is extracted as a headword.

Pattern 4: If after the question pronoun of
"چه/che/What"
the completed verb (has all inflected) is appeared then, the purpose of the question lies inside the question word and as such the descriptive level type is delivered for headword.

Pattern 5: If after the question pronoun of
"چقدر/cheghadr/How many or How much"
the linking verb is appeared then, the first noun in the phrase is extracted as a headword.

Pattern 6: If after the question pronoun of
"چقدر/cheghadr/How many or How much"
the completed verb is appeared then, the first noun in the phrase is extracted as a headword.

• Semantic Words Feature

In order to extract semantic feature for each level some groups of words are defined, each level is marked with its corresponding name. If the question's headword was in that group, the corresponding level name is added to the feature vector.

For example, if the question includes any of the words such as "day, month, year, week, hour, decades" the level name of this group is called the "date" and is added to the feature vector of that questions.

3.3 Feature Selection

After extracting all the features, the impact of every feature on the system performance was evaluated and a subset of the features that has the best precision in comparison to others is selected.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.4 Classification

The feature space is displayed as a vector so that each vector represents a question. Vector elements are the features extracted from the question. The last two columns of each feature vector are used to show the labels of coarse and fine grained classes corresponding to each question. The created set is the final dataset which is appropriate for machine learning model. The classification simulation was conducted using RapidMiner software. It was performed on this dataset using naive *Bayes*, KNN and SVM machine learning algorithms and decision tree with cross-validation evaluation criteria of 10.

Since the type of the extracted answer from the system is objective oriented, the output of this system actually shows the coarse and fine class of question's objective. Ultimately, the output of this stage would allow the questions processing module in the Q/A system to deliver a set of idioms and inquiry vocabularies into the document processing module which are used for information retrieval and finding the accurate answers to the questions.

```
Algorithm
Require: Question q
Ensure: Question head word
If q.type== کجا |چطور|چرا|چگونه then
        Return null
End if
If q.type == کدام && word following q.type is verb then
        Return the first word whose tag starts with NN
        else
        Return the word following word " کدام"
End if
 If q.type =- چه && q matches pattern1 then
      Return word following word "چه"
   Else
If q.type == چه && q matches pattern2||pattern3 then
      Return the first word whose tag starts with NN
        Else
        If q.type == چه &&q matches pattern4
        Return null
End if
If q.type == چند && word following q.type is verb then
        Return the first word whose tag starts with NN
        Else Return the word following word " چند"
        End if
If q.type== کیست| چیست then
      Return the first word whose tag starts with NN
End if
```

Figure 2. Headword Extraction Algorithm

4. SYSTEM EVALUATION

In order to test and evaluate the proposed algorithm the RapidMiner software was used. Typically, the performance of a question classifier is measured by calculating the accuracy of that classifier on a particular test set. The accuracy in question classification is defined:

$$Accuracy = \frac{no.\ of\ classified\ samples}{\text{Total no. of Tested samples}} \quad (1)$$

There are also two class-specific performance metrics: precision and recall, which can be used in question classification problem. The precision and recall of a classifier on a particular class c are defined as follow:

$$Precision[c] = \frac{\text{no. of Samples Correctly Classified as c}}{\text{no. of Samples Classified as c}} \quad (2)$$

$$Recall[c] = \frac{\text{no. of Samples Correctly Classified as c}}{\text{Total no. of Samples in Class c}} \quad (3)$$

The created features matrix in the previous step was introduced as input data to the algorithm. In this study, two different tests have been carried out.

The first test, the proposed approach with different classification algorithms, has been conducted for both coarse and fine grained class and the algorithm with best performance was used for subsequent experiments. In the next test, the features impact on the performance of both coarse and fine grained classes is studied. Finally, the proposed approach was compared with two works done on the Persian language.

In order to test the proposed approach with different classification algorithms, the naive *Bayes* from Bayesian classifier algorithms categories, KNN from the nearest neighbor's categories, a library for support vector machines (LIBSVM) from SVM classifier categories and decision tree from the tree classifier categories were selected. For SVM categories it should be noted that it is usually used for two class problems while our problem is considered as multiple class. Therefore, LIBSVM algorithm is selected as a representative for SVM class because it is the only one in SVM class which has the ability to classify the multiple problems internally.

By adjusting the algorithm's parameters, the achieved results on both coarse and fine grained classes are shown in Figures 3 and 4 respectively.

Journal of Theoretical and Applied Information Technology
15th November 2018. Vol.96. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

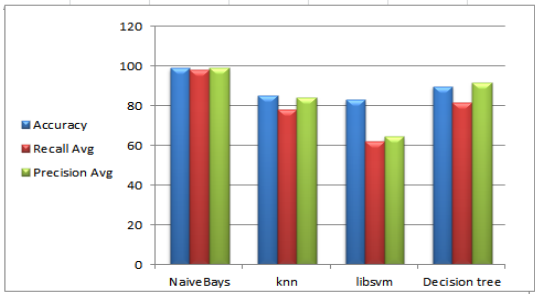

Figure 3. Proposed Approach Performance with Different Algorithms in Coarse Classification

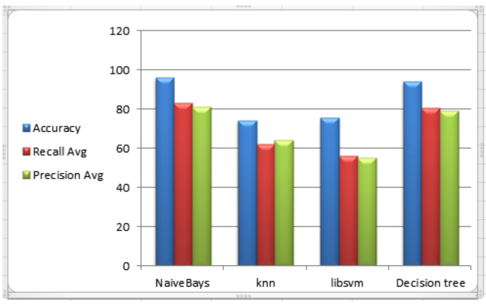

Figure 4. Proposed Approach Performance with Different Algorithms in Fine Classification

This experiment was conducted with 301 Persian questions in nine deferent question models. The minimum length of the question is 2 words and maximum is 15, and the highest frequency of questions are 6-7 words. These sentences include 106 word headwords and 39 semantic words. The distribution of coarse grained classes are presented in Table 2.

Table 2: Data Distribution in Coarse Classification

Index	Nominal Value	Absolute	Fraction
1	Entity	102	0.339
2	Number	84	0.279
3	Description	53	0.176
4	Location	42	0.140
5	Human	14	0.047
6	Abbreviation	6	0.020

As can be seen from Figures 3 and 4, for both coarse and fine grained classes According to the test conditions, the naive Bayes has outperformed other algorithms.

In addition to the common features used in other works, in this paper the headword feature and the related semantic feature on Persian questions are introduced and the impact of these two features are also investigated in this section. For this purpose, in one time both introduced features are removed and the system performance has been tested and it was observed that for coarse grained class the classification accuracy was reduced from 99% to 93%. By including the headword feature the classification accuracy was increased to 97.67% and after inclusion of second feature i.e. the related semantic feature, the accuracy was again increased to 99%.

What is interesting here is that any of these two introduced features separately has very considerable effect on all three criteria. The highest classification accuracy and recall is related to a time when only semantic features are used and the highest accuracy is achieved when all the features are used. Table 3 shows the achieved results for all three classification criteria.

Table 3: Effect of Headword and Related Semantic word Features in Coarse Classification

	Accuracy	Avg.Precision	Avg.Recall
All features except headword & semantic word	93.02%	86.43%	94.06%
All features except semantic word	97.67%	97.08%	97.44%
All features	99%	99.13%	98.25%
Feature headword & semantic word	99%	95.14%	98.28%
Feature headword	98.34%	93.75%	98.09%
Feature semantic word	99.34%	97.23%	99.64%

The same test was also conducted on fine grained classes. The results show that without having these two introduced features the system has very poor performance. By including the headword feature the classification accuracy was increased to 81% and after inclusion of second feature, the accuracy was increased to 96%. The best system performance is related to a time when the introduced features are used. Table 4 shows the achieved results for the effect of features on fine grained classes.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4:Effect of Headword And Related Semantic word Features in fine Classification

	Accuracy	Avg.Precision	Avg. Recall
All features except headword & semantic word	32.89%	29.94%	24.87%
All features except semantic word	81.40%	70.51%	68.40%
All features except headword	95.67%	82%	84.22%
All features	96.01%	81.4%	83.14%
Feature headword & semantic word	99%	90.86%	92.86%
Feature semantic word	98.67%	90.82%	92.22%
Feature headword	78.74%	76%	71%

For Persian questions two researches have been carried out, one for General Persian questions that its classification is similar to the utilized classification in this paper. The second one has been carried out on the questions appeared in Quran which its classification is different and for comparison that dataset is utilized.

It should be noted, for related semantic words feature, since the dataset has been changed, the dataset includes related semantic words with subset of new Quran dataset. The classification achieved comparison results between proposed algorithm using Naive Bayes technique and the conducted works in Farhoodi [16] and company are presented in Tables 5 and 6. Table 7 demonstrates the comparison results between proposed algorithm and the conducted work by Mollaei et.al [18].

The achieved results presented in Tables 5, 6 and 7shows that our method has outperformed other mentioned methods in terms of accuracy for both coarse and fine grained classes.

Table 5. Comparison of Our Proposed Approach and provided Approach by Farhoodi [16] for Coarse Classification (Quran Dataset)

Approach	Accuracy	*Avg.Precision*	*Avg. Recall*	F measure
proposed approach	97.25%	87.36%	88.99%	88.16%
Farhoodi [16]	76.55%	76.56%	80.55%	76.07%

Table 6. Comparison of Our Proposed Approach and provided Approach by Farhoodi[16] for Fine Classification (Quran Dataset)

Approach	Accuracy	Avg.Precision	Avg. Recall	F measure
proposed approach	96.44%	80.44%	83.89%	82.12%
Farhoodi [16]	72.17%	75.92%	71.25%	71.28%

Table 7. Comparison of Proposed Approach and Provided Approach by Mollaei [18]

	Coarse classification accuracy	Fine classification accuracy
proposed approach	99.34%	99%
Mollaei [18]	87.20%	85.31%

5. CONCLUSION

The obtained results on the effect of two introduced features i.e. question's headword features and in our main dataset (including general questions) showed that both these features are effective independently to improve coarse and fine grained classes. But if the semantic feature is available then the effect of headword feature to improve the coarse grained class is minor such that for this class the highest classification accuracy was achieved when only the related semantic words feature was utilized.

It is important to say that the effect of introduced features on various datasets is different. For example, for Quran questions set both features contributed to improve the accuracy and the highest accuracy was achieved once both features were utilized.

Moreover, as another important hint, the related semantic word can be extracted by the help of headword. In fact, in case of the availability of headword in the question the semantic subclass related to extracted headword is delivered as a feature of semantic word related to that question. In other word, the more accurate headword extraction will cause the higher accuracy in related semantic word extraction.

Generally, it can be reported that inclusion of these two presented features have very effective role on improving the classification accuracy for both coarse and fine grained classes. It can be concluded, without the existence of these features the classification performance was very poor for both classes especially for fine grained class. Therefore, it cannot say that the headword feature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

is has an effective role in classification accuracy of coarse grained class and the related semantic words feature has an effective role in classification accuracy of fine grained class. But it can say that both these features have contribution on improving the performance of both classes and the role of these features is very considerable to improve the performance of fine grained class.

Given that the extraction of headword is dependent on the syntactic structure of words, so whatever PoS label is accurate, the extraction of the headword will be more accurate. Unfortunately, Persian text pre-processing tools are limited and do away with high precision. In the Persian language the words may be written in multiple writing fashions. In this case, a question included with this word has two different written forms and the delivered headword for the same question in each case is different that this is a weakness and this problem should be eliminated in the future with improved text pre-processing schemes.

However, in this study with regard to the related semantic words to a significant extent these weaknesses have been covered. This means that both extracted headwords from these questions are related to a semantic group and delivers a same subclass.

In this study, the standard Persian questions were used and due to the limited number of questions, a number of standard questions from TREC set have been translated into Persian. However, some subclasses doesn't have question or they have limited number of questions. Thus, in the future it is needed to create a comprehensive Persian question dataset.

In this research questions starting with word "آيا" which have "yes/no" answers are not covered in our proposed classification algorithm. Therefore, in the future these types of questions and their relevant answers must be included in the classification. Moreover, there are several question phrases in Persian which they don't have the question word so they also must be taken into our consideration for the future work.

REFERENCES

[1] Kwok CCT, Etzioni O, Weld DS ,"Scaling question answering to the web", In: WWW '01: proceedings of the 10th international conference on WorldWideWeb, ACM New York, NY, USA, pp. 150–161,2001.

[2] N. Laokulrat, "A survey on question classification techniques for question answering", KMITL Information Technology Journal, vol. 2, no. 1,2013.

[3] Dell Zhang, Wee Sun Lee," Question Classification using Support Vector Machines", Proceedings of the 26th annual international ACM dl.acm.org., pp.26-32,2003.

[4] Elena Verdú , et al, "A genetic fuzzy expert system for automatic question classification in a competitive learning environment" , Elsevier,vol.39,pp.7471-7478,2012.

[5] Li liu,et al,"Chinese Question Classification based on Question property kernel",international learning and Cybernetics,vol.5,pp.713-720,2014.

[6] Li, Xin, Dan Roth, "Learning question classifiers", In Proceedings of the 19th international conference on Computational linguistics, Vol. 1, pp. 1-7,2002.

[7] Gupta, P, Gupta,Vishal, "A Survey of Text Question Answering Techniques", International Journal of Computer Applications, vol.53,No. 4,pp.1-8,2012.

[8] Juan Le,et al," Question Classification Based on Fine-Grained PoS Annotation of Nouns and Interrogative Pronouns", Trends in Artificial Intelligence, Springer, vol.8862,pp.680-693,2014.

[9] Rishika Yadav , Megha Mishra, "Question Classification Using Naïve Bayes Machine Learning Approach" , International Journal of Engineering and Innovative Technology (IJEIT),vol.2,pp.291-294,2013.

[10] Luka Bradesko,et al, " Contextualized Question Answering", Journal of Computing and Information Technology ,vol.18, No.4, pp.325-332,2010.

[11] Hamid Rastegari and Siti Mariyam Shamsuddin," Web Search Personalization Based on Browsing *History by Artificial Immune System*",Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 3, November 2010.

[12] Phil Blunsom, Krystle Kocik, and James R. Curran. "Question classification with log-linear models", In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR, ACM, pp.615–616, New York, NY, USA,2006.

[13] Krishnan, V., Das, S. & Chakrabarti, S., "Enhanced Answer Type Inference from Questions Using Sequential Models", In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada,Association for Computational Linguistics, pp. 315-322, 2005.

[14] Babak Loni, "Enhanced Question Classification with Optimal Combination of Features", Department of Media and Knowledge Engineering, Delft University of Technology, Master Thesis,2010-2011.

[15] Shima Rakian and Faramarz Safi Esfahani and Hamid Rastegari," A Persian Fuzzy Plagiarism Detection Approach",Journal of Information Systems and Telecommunication, Vol. 3, No. 3, July-September 2015.

[16] Sherkat, ehsan, Farhoodi,mojgan, "A Hybrid Approach for Question Classification in Persian Automatic Question Answering Systems", 4th International eConference on Computer and Knowledge Engineering (ICCKE), 2014 IEEE, pp.279-284, 2014 .

[17] O. Kolomiyets, and M.F. Moens, "A survey on question answering technology from an information retrieval perspective", Information science, vol. 181, no. 24, pp. 5412-5434,Dec. 2011.

[18] Ali Mollaei, Rahati quchani Saeed, Estaji Azam," Question classification in Persian language based on conditional random fields", Computer and Knowledge Engineering (ICCKE), 2nd International eConference on, IEEE, 2012.