Detecting Question Intention Using a K-Nearest Neighbor Based Approach

Alaa Mohasseb[1], Mohamed Bader-El-Den[1], Mihaela Cocea[1]

1. School of Computing
University of Portsmouth, Portsmouth, UK
Email: {alaa.mohasseb, mohamed.bader, mihaela.cocea}@port.ac.uk

Abstract. The usage of question answering systems is increasing daily. People constantly use question answering systems in order to find the right answer for different kinds of information, but the abundance of available data has made the process of obtaining relevant information challenging in terms of processing and analyzing it. Many questions classification techniques have been proposed with the aim of helping in understanding the actual intent of the user's question. In this research, we have categorized different question types through introducing question type syntactical patterns for detecting question intention. In addition, a k-nearest neighbor based approach has been developed for question classification. Experiments show that our approach has a good level of accuracy in identifying different question types.

Keywords: Natural Language Processing, Question Classification, Machine Learning, Text Mining, Information Retrieval

1 Introduction

The usage of question answering systems is increasing daily, people make frequent use of question answering systems in order to find the right answer for different kinds of information. The goal of question classification process is to accurately assign labels to questions based on expected answer type.

Many questions classification techniques have been proposed with the aim of helping in understanding the actual intent of the user's question but the abundance of available data has made the process of obtaining relevant information challenging in terms of processing and analyzing it.

Recent studies classified different type of questions by using different machine learning algorithms such as Support Vector Machine (SVM) [8], [3], [6], [17]. Other works like [19] and [9] used SVM in addition to other machine learning algorithms such as Naive Bayes, Nearest Neighbors and Decision Tree. Moreover, Neural Networks has been used as the machine learning algorithm in other works [15] and [16].

Furthermore, other methods such as features selection have been applied to obtain an accurate question classifier [6], [19], [7], [18], [9] used bag-of-words,

Other works like [18] used semantic and syntactic features, Moreover, [6] used uni-gram and word shape feature.

In this paper, we propose a method that automatically identifies and classifies users' questions intention using a k-nearest neighbor based approach based on the syntactical pattern of each type of question. In particular, we develop a framework which was adapted from [11] and [10] to test the performance of the proposed method. Experimental results show that our solution leads to accurate identification of different question types.

The rest of the paper is organized as follows: Section 2 outlines previous work on question classification. Section 3 provides a detailed description of the proposed question classification framework. Section 4 reports experimental results. Finally, Section 5 concludes the paper and outlines future work.

2 Related Work

In this section we outline previous work on questions classification methods and machine learning algorithms.

Recent studies classified different type of questions by using different machine learning algorithms. In [8] a statistical classifier has been proposed which is based on SVM and uses prior knowledge about correlations between question words and types in order to learn question word specific classifiers. They have stated that under such a statistical framework, any data set, question ontology, or set of features can be used.

Other works like [19] and [9] used SVM in addition to other machine learning algorithms. [9] proposed an approach for question classification through using machine learning. In this work three different classifiers were used, which are; Nearest Neighbors (NN), Nave Bayes (NB), and SVM using two kinds of features: bag-of-words and bag-of n grams. In order to train the learning algorithm, a set of lexical, syntactic, and semantic features were used, among which are the question headword and hypernym. Similarly, in [19] five machine learning algorithms were used, which are; NN, NB, Decision Tree (DT), Sparse Network of Winnows (SNoW), and SVM using two kinds of features: bag-of-words and bag-of-ngrams.

In addition, authors in [17] proposed a method of using a feature selection algorithm to determine appropriate features corresponding to different question types. Moreover, they design a new type of features, which is based on question patterns then applied a feature selection algorithm to determine the most appropriate feature set for each type of questions. The proposed approach was tested on the benchmark dataset TREC, using SVM for the classification algorithm.

SVM were also used in [3] for the classification of open-ended questions. They have stated that SVM could be trained to recognize the occurrence of certain keywords or phrases in a question class and then, based on the recurrence of these same keywords, be able to correctly identify a question as belonging to that class.

Another classification has been proposed in [4] using SVM. According to authors in this work enormous amount of time is required to create a rich collection

of patterns and keywords for a good coverage of questions in an open-domain application, so they have used support vector machines for question classification. The goal is to replace the regular expression based classifier with a classifier that learns from a set of labeled questions and represented the questions as frequency weighted vectors of salient terms.

Moreover, works like [15] and [16] used Neural Networks as the machine learning algorithm. [15] proposed a neural network for question answering system, they have stated that the proposed network can process many complicated sentences and can be used as an associative memory and a question-answering system. In addition, the proposed network is composed of three layers and one network which are, Sentence Layer, Knowledge Layer, Deep Case Layer and Dictionary Network. The input sentences are divided into knowledge units and stored in the Knowledge Layer.

The proposed approach in [16] formulates the task as two machine learning problems which are, detecting the entities in the question, and classifying the question as one of the relation types in the knowledge base. Based on this assumption of the structure, this approach trains two recurrent neural networks and outperform state of the art by significant margins relative improvement.

Furthermore, other studies classified different type of question using different features selection like bag-of-words, semantic and syntactic features, and unigram and word shape feature.

Authors in [6] proposed head word feature which used two approaches to augment semantic features of such head words using WordNet. In addition, other standard features were augmented as well such as wh-word, unigram feature, and word shape feature.

In [18] a machine learning-based question-answering framework has been proposed, which integrates a question classifier with a simple document/passage retrievers, and proposed context-ranking models. This method provides flexible features to learners, such as word forms, syntactic features, and semantic word features. In addition, The proposed context-ranking model, which is based on the sequential labeling of tasks, this model combines rich features to predict whether the input passage is relevant to the question type.

Finally, works in [7] used machine learning approaches, namely, different classifiers and multiple classifier combination method by using composite statistic and rule classifiers, and by introducing dependency structure from Minipar and linguistic knowledge from Wordnet into question representation, in addition, features like the Dependency Structure, Wordnet Synsets, Bag-of-Word, and Bi-gram were used. Also a number of kernel functions were analysed and the influence of different ways of classifier combination, such as Voting, AdaBoost, ANN and TBL, on the precision of question classification.

3 Proposed Approach

In this section we introduce a K-nearest neighbor based approach using domain specific syntax information for question classification. The framework mainly

relies on Question Type Syntactical Patterns. In the first part of this section question types that have been used are explained in detail; in the second section we introduce our syntactical patterns and in the third part we explain the structure of the proposed approach.

3.1 Question Types

Questions could be classified according to their intent into six categories; Factoid, Choice, Causal, Confirmation, Hypothetical and List; each of these types have their own structure and characteristics [12].

1. Factoid: this type begins with a question word such as *What, Where, Why, Who, Whose, When, Which*, as well as *How, how many, how often, how far, how much, how long, how old* and any kind of information is expected as an answer. For example *"what is a good blood pressure"*.
2. Choice: this type of question offers choices in the question. The question contains two (or more) presented options. These options are connected using the conjunction "OR". For example *"Which is better iphone or samsung? and why?"*.
3. Causal: starts with How or Why and requires explanation. For Example, *"why do earthquakes occur at destructive plate margins?"*.
4. Confirmation: this type of question begins with an auxiliary verb or linking verb for example *"is Frankfurt a city in Germany?"*, in addition, the question could start with negative auxiliary verb or linking verb. For example, *"wasn't Thomas Edison born in new jersey"*. The expected answer for this type of question is either Yes or No.
5. Hypothetical: hypothetical questions are asked to have a general idea of a certain situation. It is mainly What would you do if / What would happen if" type of questions. For example *"what would you do if someone had a stroke?"*.
6. List: plural terms are a highly reliable indicator of this question, this type requires a list of entities or facts in answers. For example *"What countries are in Africa?"*.

3.2 Question Types Syntactical Patterns

The proposed framework mainly relies on the question types and the characteristics of each type discussed in Section 3.1. Using these characteristics we propose the formulation of syntactical patterns for each question; thus, these give us domain-specific information. Each syntactical pattern is composed of a sequence of term categories. These categories of terms are described below. For the purpose of constructing Question Type Syntactical Patterns, a random set of $3,000$ questions has been selected from Yahoo Non-Factoid Question Dataset[1] and TREC 2007 Question Answering Data [2].

[1] `https://ciir.cs.umass.edu/downloads/nfL6/`
[2] `http://trec.nist.gov/data/qa/t2007_qadata.html`

The categorization of terms in our solution is mainly based on the seven major word classes in English: Verb (V), Noun (N), Determiner (D), Adjective (Adj), Adverb (Adv), Preposition (P) and Conjunction (Conj). In addition to that, we added a category for question words that contains the six main question words (QW): how, who, when, where, what and which. Some word classes like Nouns consists of subclasses, such as Common Nouns (CN), Proper Nouns (PN), Pronouns (Pron) and Numeral Nouns (NN). Also, the Verb class has subclasses, such as Action Verbs (AV), linking Verbs (LV) and Auxiliary Verbs (AuxV).

Furthermore, the syntactical patterns of each question types have been identified by tagging each term in the question to one of the main word classes mentioned above, and then a further tagging is done to assign each term in the question to one of the domain specific term categories. For example, in the question *"who is Frida Kahlo"*, the terms will be tagged as follows: (a) *"who"* will be tagged to *"QW"*, (b) *"is"* is tagged to *"LV"* and (c) *"Frida Kahlo"* is tagged to *"PN"*.

Finally, after each term is tagged to one of the word classes, it will be tagged to the domain specific term category; the proposed categories are derived from the following topics;

1. *Health:* which includes specific terms related to health, medicine, beauty.
2. *Sports:* includes terms related to game and recreation, sports events, sports.
3. *Arts and entertainment:* consists of terms related to Entertainment, Celebrities Name, lyrics, Movies, Books, Authors.
4. *Food and drinks:* includes terms related to foods, drinks, recipe.
5. *Animals:* consists of terms related to Pets, wild animals.
6. *Science and math:* which includes specific terms related to Science, math.
7. *Technology and internet:* consists of terms related to Software and Applications, Site, Website, URL, Database and Servers.
8. *Society and culture:* includes terms related to Environment, Holidays, Months, history, political, Relationships, Family.
9. *News and events:* includes terms related to Newspapers, Magazines, Documents, events.
10. *Job, Education and Reference:* includes terms related to Careers, Institutions, Associations, Clubs, Parties, Foundations and Organizations.
11. *Business and Finance:* includes terms related to Money, company,products, Economy.
12. *Travel and places:* which includes specific terms related Geographical Areas, Transportation, Places and Buildings, Countries.

These categories help in the identification of the main topics that are found in most question answering systems. Terms categories have been created for the purpose of identifying the different type of questions. These terms have been constructed after the analysis of different datasets. For example, *"QW"* will be tagged to "Question Word Who" (QW_{Who}); *"LV"* will not be tagged to any further categories and *"PN"* will be tagged to "Proper Noun Celebrity" ($PN_{Celebrity}$). This step is executed by using a database that contains more than 10,000 terms [13].

3.3 Framework

To investigate the impact of using the domain specific syntax information on the classification performance, the following framework, shown in Figure 1, has been developed. The proposed framework involves automatic identification and classification of user's questions using KNN approach which is based on the patterns described in the previous section. We illustrate the framework by using the following examples of question: *"is mercury a metal"* and *"what are the symptoms of diabetes"*.

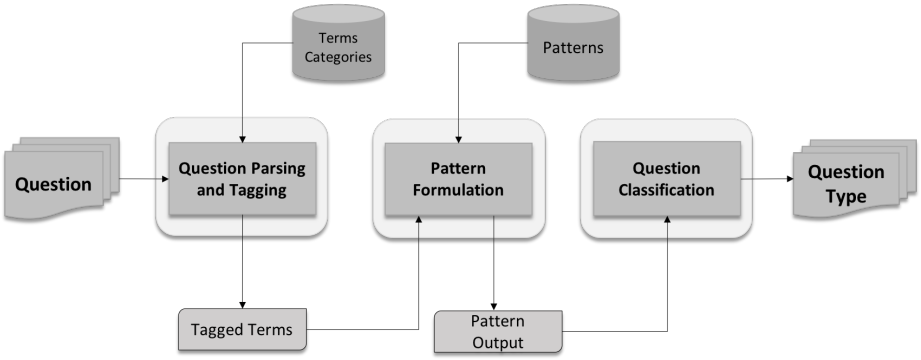

Fig. 1. Question Classification Framework

1. Question Parsing and Tagging:
   This step is mainly responsible for extracting users question terms. The system simply takes the question and parses to tag each term in the question to its terms' category.
   Question 1: is mercury a metal
   Terms extracted: is, mercury, a, metal

   Question 2: what are the symptoms of diabetes
   Terms extracted: what, are, the, symptoms, of, diabetes

   After parsing, each term in the question will be tagged to one of the terms category.
   The final tagging will be:

   Question 1 Terms Tagging:
   is=LV, *mercury=* $PN_{Science}$, *a= D, metal=* $CN_{Singular}$, where $PN_{Science}$ is Proper Noun Science and $CN_{Singular}$ is Common Noun Singular

   Question 2 Terms Tagging:

what= QW_{What}, *are*=LV, *the*= D, *symptoms*=CN_{Plural}, *of*= P, *diabetes*=CN_{Health}, where CN_{Plural} is Common Noun Plural and CN_{Health} Common Noun Health.

2. Pattern Formulation
   In this phase after tagging each term in the question, the pattern is formulated, as illustrated below for the two previously introduced questions.

   Question 1 Pattern: $LV + PN_{Science} + D + CN_{Singular}$

   Question 2 Pattern: $QW_{What} + LV + D + CN_{Plural} + P + CN_{Health}$

3. Question Classification
   In this step the system first attempts to match the question with the most appropriate Question Type Pattern to determine the Question type. For the given examples.

   Question 1 Type: Confirmation

   Question 2 Type: List

   Second, the system will match the question with the most appropriate Domain Category to determine the Question Domain Specific. This step is done by matching each words in the question with one of the Domain Specific Terms Categories 3.2 and the domain is selected based on the number of words related to a domain in each question. For the given examples.

   Question 1 Domain Category: Science and math

   Question 2 Domain Category: Health

   This will result in the final classification of each question in which *Question 1* will be classified to $Confirmation\_Scienceandmath$ and *Question 2* will be classified to $List\_Health$

4 Experimental Evaluation

To test the accuracy of our proposed approach 1,160 questions were randomly selected from Yahoo Non-Factoid Question and TREC 2007 Question Answering Data. Their distribution is given in Table 1.

K-Nearest Neighbor (KNN) was used as the machine learning algorithm for the automatic classification. The classification accuracy is obtained by using the implementation of the above algorithm from the Weka software [5]. The effectiveness of the classification was evaluated based on Precision, Recall and F-Measure, i.e. typical metrics for the evaluation of classifiers, using 10-fold

Table 1. Data distribution

Question type	**Total**
Causal	31
Choice	12
Confirmation	321
Factoid	688
Hypothetical	7
List	101

cross-validation and value of K=1. Furthermore, a comparison was done using different values of K to evaluated the classification accuracy when increasing the value of K from K=1 to 10.

Table 2 presents the classification performance details (Precision, Recall and F-Measure) of the KNN classifier. Results show that KNN identified correctly (i.e. Recall) 82.8% of the questions when the value of K=1.

KNN classified correctly 93.1% (Recall) of the confirmation questions and 92.2% of the factoid questions, on the other hand, classification accuracy (Recall) for causal, choice, hypothetical and list questions were lower. KNN could correctly classified 32.3% of the causal questions, 25% of the choice questions, 28.6% of the hypothetical questions and 12.9% of the list questions.

Table 2. KNN classifier performance with value of K=1

Question Types	**Precision**	**Recall**	**F-Measure**
Causal	0.714	0.323	0.444
Choice	0.333	0.250	0.286
Confirmation	0.849	0.931	0.889
Factoid	0.853	0.922	0.886
Hypothetical	0.667	0.286	0.400
List	0.333	0.129	0.186
Overall	0.797	**0.828**	0.805

Figure 2 shows the impact of increasing the value of K, There is a marginal increase in the accuracy between k=1, 2 and 3 in which K=3 has the highest accuracy with 83.7% shown in table 3, however, in terms of Recall for certain question categories such as causal, choice, hypothetical and list K=1 performed better. In addition, when K=4, 5 and 6 the accuracy slightly decreased while when the value of K= 6 and 7 the accuracy increased again but decreased with K=9 and 10.

These results show that KNN deals well with confirmation and factoid questions. In addition, KNN could not distinguish between causal, choice, hypothetical and list type of questions and incorrectly classified most of them as confirmation and factoid questions. As shown in table 1 the question dataset suffers from imbalance between the labels, as the number of the instances that

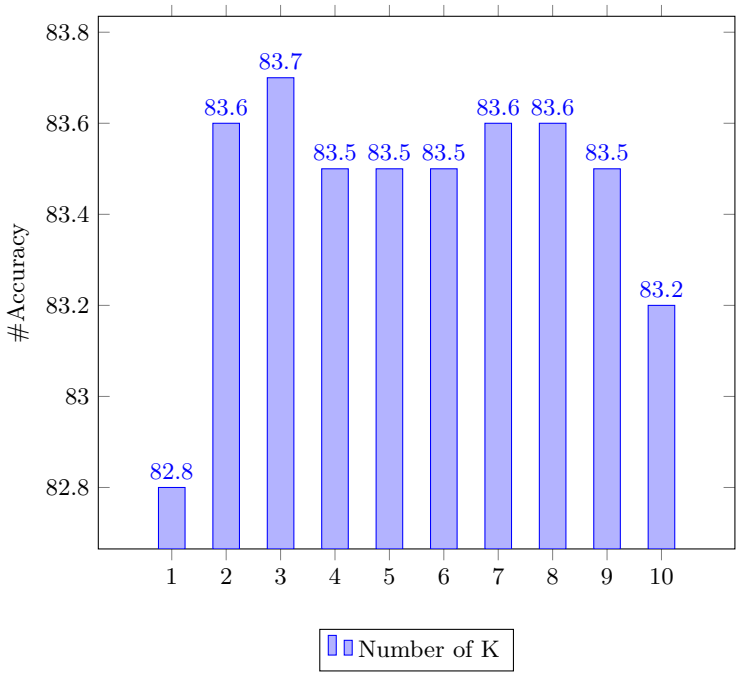

Fig. 2. Comparison of using different value of K

belongs to one more classes is outnumbered by the instances that belong to other classes. This is know by the class imbalance problem and it normally hinders the classifiers ability in correctly predicting the instances that belongs to the minority class [14]. However, the proposed algorithms has been able to perform reasonably well over the minority classes as shown in table 2 and table 3.

Table 3. KNN classifier performance with value of K=3

Question Types	**Precision**	**Recall**	**F-Measure**
Causal	0.714	0.161	0.263
Choice	1.000	0.083	0.154
Confirmation	0.870	0.938	0.903
Factoid	0.831	0.956	0.889
Hypothetical	1.000	0.143	0.250
List	0.385	0.050	0.088
Overall	0.802	**0.837**	0.795

Overall, the results validate that questions Type Syntactical Patterns is an effective method for question classification as well as for the distinction between different question types.

5 Conclusions

In this paper, we have proposed a method that automatically identifies and classifies different question types by using a domain specific syntax information, which is based on the syntactical pattern of each types of question. In particular, we developed a framework to test the performance of the proposed method and used a machine learning algorithm (KNN) to build a model for the identification of user's question type. The experiment shows that our solution led to a good performance in classifying questions.

As future work, we aim at examining and analyzing more questions from different data-sets and extending the analysis of the different types of questions. As mention earlier, the questions dataset suffers from imbalance between the label, we aim to investigate ensemble learning and other methods to deal with the class imbalance problem [2], [1]. We are also planning to test other machine learning algorithms to classify the questions. In addition, we will test our data-set using other classification frameworks in order to evaluate our method and be able compare it with different approaches.

References

1. Bader-El-Den, M.: Self-adaptive heterogeneous random forest. In: Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on. pp. 640–646. IEEE (2014)
2. Bader-El-Den, M., Teitei, E., Adda, M.: Hierarchical classification for dealing with the class imbalance problem. In: Neural Networks (IJCNN), 2016 International Joint Conference on. pp. 3584–3591. IEEE (2016)
3. Bullington, J., Endres, I., Rahman, M.: Open ended question classification using support vector machines. MAICS 2007 (2007)
4. Hacioglu, K., Ward, W.: Question classification with support vector machines and error correcting codes. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2. pp. 28–30. Association for Computational Linguistics (2003)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
6. Huang, Z., Thint, M., Qin, Z.: Question classification using head words and their hypernyms. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 927–936. Association for Computational Linguistics (2008)
7. Li, X., Huang, X.J., WU, L.d.: Question classification using multiple classifiers. In: Proceedings of the 5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network (2005)
8. Metzler, D., Croft, W.B.: Analysis of statistical question classification for fact-based questions. Information Retrieval 8(3), 481–504 (2005)
9. Mishra, M., Mishra, V.K., Sharma, H.: Question classification using semantic, syntactic and lexical features. International Journal of Web & Semantic Technology 4(3), 39 (2013)

10. Mohasseb, A., Bader-El-Den, M., Liu, H., Cocea, M.: Domain specific syntax based approach for text classification in machine learning context. In: 2017 International Conference on Machine Learning and Cybernetics (ICMLC). vol. 2, pp. 658–663. IEEE Systems, Man and Cybernetics (2017)
11. Mohasseb, A., Bader-El-Den, M., Kanavos, A., Cocea, M.: Web queries classification based on the syntactical patterns of search types. In: International Conference on Speech and Computer. pp. 809–819. Springer (2017)
12. Mohasseb, A., Bader-El-Den, M., Cocea, M.: Question categorization and classification using grammar based approach. Information processing and management p. Under Review (2018)
13. Mohasseb, A., El-Sayed, M., Mahar, K.: Automated identification of web queries using search type patterns. In: WEBIST (2). pp. 295–304 (2014)
14. Perry, T., Bader-El-Den, M., Cooper, S.: Imbalanced classification using genetically optimized cost sensitive classifiers. In: Evolutionary Computation (CEC), 2015 IEEE Congress on. pp. 680–687. IEEE (2015)
15. Sagara, T., Hagiwara, M.: Natural language neural network and its application to question-answering system. Neurocomputing 142, 201–208 (2014)
16. Ture, F., Jojic, O.: Simple and effective question answering with recurrent neural networks. arXiv preprint arXiv:1606.05029 (2016)
17. Van-Tu, N., Anh-Cuong, L.: Improving question classification by feature extraction and selection. Indian Journal of Science and Technology 9(17) (2016)
18. Yen, S.J., Wu, Y.C., Yang, J.C., Lee, Y.S., Lee, C.J., Liu, J.J.: A support vector machine-based context-ranking model for question answering. Information Sciences 224, 77–87 (2013)
19. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 26–32. ACM (2003)