Journal of Theoretical and Applied Information Technology
31st January 2015. Vol.71 No.3

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

QUESTION CLASSIFICATION USING STATISTICAL APPROACH: A COMPLETE REVIEW

[1]ANBUSELVAN SANGODIAH, [2]MANORANJITHAM MUNIANDY, [3]LIM EAN HENG

[1, 3]Department of Information System, Faculty of Information and Communication Technology, University Tunku Abdul Rahman

[2]Department of Computer Science, Faculty of Information and Communication Technology, University Tunku Abdul Rahman

[3]Department of Information System, Faculty of Information and Communication Technology, University Tunku Abdul Rahman

E-mail: [1]anbuselvan@utar.edu.my, [2]manoranm@utar.edu.my, [3]ehlim@utar.edu.my

ABSTRACT

In recent days, question classification is growing in popularity as it has an important role in question answering systems, information retrieval and it can be used in a wide range of other domains. The main aim of question classification is to accurately assign labels to questions based on expected answer type. Past research works have relied on matching questions against hand-crafted rules. However, rules require enormous effort to create and often suffer from being too specific. A great deal of current research works on question classification is based on statistical approach to overcome these issues by employing machine learning techniques such as Support Vector Machine and Artificial Neural Network. This paper presents an updated literature survey of current methods or approaches for question classification in the areas of question answering systems, information retrieval and educational environment. Question classification involving other languages besides English has also been examined.

Keywords: *Question Classification, Machine Learning, Semantic Features, Syntactic Features, Support Vector Machine*

1. INTRODUCTION

In general, question classification is the process by which a system analyzes a question and labels the question based on its expected answer type [1]. Given a finite set of possible expected answer types, known as question ontology or taxonomy or category, the goal of a question classification system is to learn a mapping from questions to answer types. Although this task may sound simple, there are many factors that determine how well such systems perform and how robust they are.

Though question classification works in the same manner as document classification the former is extremely challenging as compared to document classification in achieving a reasonable accuracy in classifying questions. This is because in question classification there exist only little information or words which may not be enough to effectively classify questions as opposed to document classification [2]. This will have impact on discriminating power in classifying questions. Nevertheless some work has been done is evident in [2], [3]. Most of these works were carried out for question answering system and information retrieval system.

Question classification systems are primarily used as components of question answering (QA) systems. QA is the task of retrieving answers to questions posed in natural language from a collection of documents, where an answer is generally a short fragment of text drawn from the corpus [1]. QA systems are a shift away from classical document retrieval towards information retrieval. This saves the user valuable time by eliminating the need to search through a long ranked list of documents for an answer to their question.

Question classification systems are not just confined to QA but information retrieval (IR) as well. It is the task of returning documents that are relevant to a particular natural language query [4]. A classic example is a query submitted by users through Google search and the query needs to be classified or labeled accurately in order to find relevant documents.

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

Question classification systems have also deeply rooted in educational environment though its inherent nature of the problem that dictates the use of question classification is different from QA and IR. Assessments in higher learning institutions or in schools usually will consist of exam questions and multiple choice questions. Cognitive measurement of learners such as Bloom taxonomy requires the questions need to be classified or labeled in accordance to the taxonomy in order to measure cognitive level of learners [5].

Question classification involving other languages besides English is also evident. Classifying questions involving Chinese language has been the focus among some researchers [3], [6], [7].

This paper is organized as follows: Sec.2 presents the use of technique or methods in question classification in QA and IR. It will mainly focus on factual type of questions and not complex questions. Sec. 3 will highlight question classification in educational environment. It will focus on imperative and factual types of questions. Sec. 4 presents question classification involving other languages apart from English where the area of application is particularly in QA. At the end of this paper, a comparison table for each section is worked out to illustrate better the differences and the similarities between one research work with another.

2. QUESTION CLASSIFICATION IN QA AND IR

Research on question classification using statistical approach has come a long way since as early as 2002. Some of the earlier question classification work includes [8] and [9], in which language models and Rappier rule learning were employed respectively. In [8], authors have proposed the use of QuASM which is semi structured metadata in question answering system. The main goal is to answer factual questions by exploiting the structure inherent in documents found on the World Wide Web (WWW) and thus improving search performance by providing users with specific answers, rather than having users scan retrieved documents for these answers. The semi structured nature of web documents offer clues to where sections begin and end, as well as their subject matter. Using these clues, it is believed that documents can be broken into smaller units that are typically homogeneous. The hypothesis is that by indexing smaller parts of documents for IR, the answer search will be more effective. The QuASM experiments also involved query classification. In order to find answers, a relationship must be established between the question being asked and possible target answers in the document. Questions need to be classified on the basis of the type of answer that is expected. These answer entities must also be recognized in retrieved documents so the answers can be located. To achieve this, regular expression and language models such as n-gram models particularly unigram and bigram have been used to estimate the probability that the answer class generated the question. Despite findings show some improvements in using QuASM in retrieving relevant answer to a query, the author did highlight that much improvement is required in order to get more promising results. However, the question classifier using bigram and unigram without tagging in QuASM experiment recorded a precision between 73% and 74% while regular expression is 59%.

In [9], authors have proposed probabilistic phrase re-ranking algorithm (PPR), which uses proximity and question type features to extend the capability of search engines to support natural language question answering. The stages in PPR are query modulation, document retrieval, passage (or sentence) retrieval, phrase (answer) extraction, and answer ranking. Before retrieving relevant documents and providing factual answers to queries by users, identification and classification of question type is important. In this research, decision rule induction using Ripper [10] and a heuristic rule-based algorithm have been used to identify question type. In the experiment, each question is represented by 13 features, 9 of which are semantic features based on WordNet where Ripper is deployed for question type identification. The second method for question type identification that has been used is heuristic algorithm. It uses POS tagger to identify question type for questions containing no wh-word and has what/which words. Results indicate that the use of both features which are proximity and question type improves the accuracy in retrieving actual answers to queries. Also in the experiment, different levels of granularity of document content have been studied to observe how effectively the chances of finding good answers from the search results can be improved. The answers of different levels are document, sentence, and phrase. In overall comparison, results show that use of sentence and phrase answer levels with features of proximity and question type give significant impact in ranking actual answers.

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

More recently, Li and Roth, [2] have developed a machine learning approach which uses the hierarchical classifier of SNoW learning architecture [11]. They have compiled the UIUC question classification dataset 1 which consists of 5500 training and 500 test questions. The questions in this dataset are collected from four sources: 4,500 English questions published by USC [12] about 500 manually constructed questions for a few rare classes, 894 TREC 8 and TREC 9 questions, and also 500 questions from TREC 10 which serve as the test dataset. All questions in the dataset have been manually labeled by them according to the coarse and fine grained categories as shown in Table 1, with coarse classes (in bold) followed by their fine class refinements. In addition, the table shows the distribution of the 500 test questions over such categories. The primitive feature types extracted for each question include lexical words, part of speech tags, chunks (non-overlapping phrases), head chunks (the first noun chunk in a question) and named entities. They achieved 78.8% accuracy for 50 fine grained classes. With a hand built dictionary of semantically related words, their system is able to reach 84.2%. The experiment uses hierarchical classifiers and despite it achieves a decent accuracy for fine grained classes, results show that there is not performance advantage for using a level of coarse classes and the semantically appealing course classes do not contribute to better performance. In addition to hierarchical classifier, flat classifier was used in classifying fine classes, as to know whether the hierarchical classifier has any advantage in performance, in addition to the advantages it might have in downstream processing and comprehensibility.

Table 1: Question Hierarchy

Class	Class	Class	Class
ABBREV	letter	individual	**NUMERIC**
exp	other	title	code
abb	plant	description	count
ENTITY	product	**LOCATION**	date
animal	religion	city	distance
body	sport	country	money
color	substance	mountain	order
creative	symbol	other	other
currency	technique	state	period
dismed	term	**DESCRIPTION**	percent
event	vehicle	definition	speed
food	word	description	temp
instrument	**HUMAN**	manner	size
lang	group	reason	weight

The UIUC dataset has laid a platform for the follow-up research. Hacioglu and Ward [13] used linear support vector machines with question word bigrams and error-correcting output but no (Named Entity) NE tagger or related word dictionary to obtain accuracy of 80.2% to 82.0%. Generally the system has been augmented with named entity tagger and SVD based transformation. Although the NE tagging has improved the performance, it has been observed that some performance loss with SVD transformation. The main contribution of this work is the avoidance of computationally expensive linguistic analysis in feature extraction process as evidenced in [2].

Finally, Zhang and Lee's [14] question classification system is based on Support Vector Machine. The system uses a tree kernel and simple syntactic structures. It is trained and tested on the same data set and question ontology used by [2]. The system achieves 90% accuracy on the course grained question types which outperforms SVM based on word and n-gram kernels, hence proves that tree kernel enables SVM to take advantage of the syntactic structure of questions. In addition to this, the authors also experimented with question classification using different machine learning algorithms based on bag-of-words and n-grams features and results show that with only surface text features the SVM outperforms other machine learning algorithms with accuracy of 79.2%for fine grained question types and 87.4% for coarse grained question types.

Most recently, Krishnan et al. [15] used a short (typically one to three words) subsequence of question tokens or better known as informer span as features for question classification. Their model can reach the accuracy of 86.2% using UIUC dataset over fine grained question categories, which is the highest reported accuracy on UIUC dataset. And for coarse grained an accuracy of 93.4% was recorded. The model is based on meta classifier using a linear SVM on the CRF output. In this work, SVM based on perfect informer was also experimented and an accuracy of 88% was recorded over fine grained question categories while for coarse grained was 94.2%. The features of perfect informers are n-gram and n-gram hypernyms.

Later Li and Roth [16] used more semantic information sources including named entities, WordNet senses, class-specific related words, and distributional similarity based categories in question classification task. Hierarchical classifier based on SYN is used in the experiment. With all these semantic features plus the syntactic ones, their model was trained on 21500 questions and was able to achieve the best accuracy of 89.3% on a test set of 1000 questions (taken from TREC 10 and TREC

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

11) for 50 fine classes and recorded 92.3% accuracy for coarse grained classes.

In contrast to Li and Roth [16] approach which makes use of a very rich feature set, the authors [24] have proposed to use a compact yet effective feature set. The authors also commented that the work by [15] which used one contiguous span of tokens which is denoted as the informer span as features, noisy information could be introduced. To address the issues, it has been proposed head word feature and present two approaches to augment semantic features of such head words using WordNet. The WordNet semantic features for head word employ direct and indirect use of hypernyms. In addition, Lesk's word sense disambiguation (WSD) algorithm is adapted and the depth of hypernym feature is optimized. Generally, in the experiment, each question is represented by five binary feature sets which are question wh-word, head word, WordNet semantic features for head, word grams and word shape feature. The five feature sets will be separately used by the classifiers to determine their individual contribution. In addition, these features are used in an incremental fashion. With further augment of other standard features such as unigrams, the work obtains accuracy of 89.2% using linear SVMs, or 89.0% using Maximum Entropy (ME) for 50 fine classes.

In order to improve the work by [2] in terms of hierarchical classifiers particularly, in [17] authors have experimented with a variety of classifiers (SVM, MaxEnt, NB, DecisionTree) for primary and secondary classification, and findings show that a mix of a Maximum Entropy course classifier with a Naive Bayes fine classifier was the best combination. Observations indicated that it was better to mix classifiers than to have the same type of classifier as both the primary and secondary. An accuracy of 80% without WordNet was recorded using features such as bigram and parse signatures which exceeds the performance achieved by [2] (78.8%) without adding related words, so the authors expect that it should be possible to exceed [2] if the work experimented more with the semantic information.

Donald Metzler et al. [1] have come up with analysis of statistical question classification for fact based questions. SVM classifier was performed across several data sets as each data set has different characteristics, such as the expressiveness of its question ontology and its source. The data sets that have been used are TREC, UIUC and MADSCI. The main aim of the work is to examine the role different syntactic and semantic features have on performance. Results show that semantic features tend to increase performance more than purely syntactic features.

In a bid to improve the work by [2], in [18] the authors have proposed to use semi-supervised learning to consume unlabeled questions with expectation to improve the performance of classifiers in the hierarchy. In this work authors modified the Tri-training [19] which is a co-training [20],[21] style algorithm, to make it more suitable for question data type. In the experiments, authors implemented different combinations of supervised, semi-supervised with hierarchical architecture which are supervised learning for all classifiers at two levels, semi-supervised learning for the coarse-grained classifier at the first level and supervised learning for other classifiers and semi-supervised learning for all classifiers. The results prove that when the size of training sets is small, supervised learning for all classifiers in the first method is better than flat classification; the second method gives the best performance while the semi-supervised learning for fine classifiers in the third method does not successfully improve their performance. The second method uses hierarchical classification with semi-supervised learning for coarse classifier and supervised learning for fine classifiers which use MEM and SVM, respectively. The results show that the improvement of semi-supervised learning does help to improve the final precision. Interestingly, when the size of training sets increase, the performance of fine classifiers with MEM is improved better than that of fine classifiers with SVM. Consequently, the overall performance of the hierarchical classifiers in which fine classifiers using MEM which is 80.4% is improved significantly with the training set of the size 4000 compared to using SVM which is 79.0%. Another effort from this work is to try to further expand nodes in the question taxonomy that consist of a large number of question classes. As clustering is an unsupervised method to group classes that are closed (in a certain distance) to each other, clustering has been used to solve the problem. The newly created taxonomy for coarse classes such as Entity and Numeric shows that the method is in a right direction. Throughout the experiment, TREC data set was used and bag-of-words (BOW) was used as features.

In contrast to Tri Thanh Nguyen et al., [18] approach which attempts to improve hierarchical classifier, Fangtao Li et al., [22] have proposed to use Long Dependency Conditional Random Fields (LDCRFs) classifier. The authors

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

claim that not only the classifier can integrate rich features but also state and transition features as well. The features that have been used in the work are words, POS, chunker, parser information, question length, named entity, noun hypernym, head verb synset and transition features. Results show that the accuracy achieved is 85.60% which outperforms the accuracy obtained in [2] and the overall precision in [18].

Though much earlier review focuses on question classification for factual questions, the work by [23] demonstrates how support vector machines (SVM) are used successfully in the classification of open-ended questions. In the experiment, bag-of-words used as features and with some pre-processing work, an average accuracy of 74.6% was recorded. A total of 11 predefined classes or labels have been used and some classes has had accuracy exceeding 80% while others even recorded an accuracy as low as 50%. The authors believe that the poor performance for this classes or labels are attributed to small training set which has had an adverse effect in discriminating power between classes.

3. QUESTION CLASSIFICATION IN EDUCATIONAL ENVIRONMENT

As much research work of question classification as may sound in QA and IR, the same in educational environment is also no exception. Using Artificial Neural Network to classify questions for e-learning was done by [25]. In this work, multiple choice questions are automatically classified into three difficulty levels which are hard, medium and easy and these serve as class labels. The work has considered five aspects in feature selections which are query-text relevance, mean term frequency, length of Q&A, term frequency and distribution of Q & A in text. Working on small data set consisting of 233 questions, the best results of F1 obtained is 78% with number of neurons is set to 3. The authors highlighted the limitations of the work are lacking of semantic analysis and working on small data set.

Anwar Ali Yahya et al., [26] have used linear Support Vector Machines to classify questions in accordance to Bloom taxonomy and though preliminary results show a satisfactory performance of SVM in terms of accuracy and precision however a poor recall and F-measure values have been reported. The work did not consider syntactic structure and semantic knowledge of questions to improve the effectiveness in classifying questions. An accuracy of 87.4% was recorded with considering only bag-of-word as features. The size of data set is 272 questions.

Similar work by [5] has used Artificial Neural Network classifier to classify questions in accordance to Bloom taxonomy. This work is very much focused on introducing three types of features sets which are whole feature set, DF feature set and CF-DF feature set. Experimental comparisons were then conducted to investigate the effectiveness of the proposed feature reduction methods in solving classification problems based on the three performance criteria of convergence time, convergence error, and classification precision. The whole feature set represents the initial feature set with a dimension of 605 features. The DF feature set reduces the dimensionality of the feature vector by removing features that have document frequency values of 1 within each category. Therefore, the dimensionality of the DF feature set was reduced from 605 to 145, for a 76.0% reduction in dimensionality. The CF-DF feature set removes the features within each sub feature set that have category frequency values of 6 and a document frequency value of one. With this feature reduction method, the dimensionality of the feature was reduced from 605 to 141, for a 76.7% reduction in dimensionality. Results show that in terms of classification precision, the whole feature set fare better which is 65.9% than the other feature sets. However, the percentage difference between the whole feature set and the DF and CF-DF feature sets was not large and is within an acceptable range, i.e., 3.06% for DF and 4.26% for CF-DF.

Rule based classifier can also be used in classifying questions in accordance to Bloom taxonomy and the work by [27] achieves it. This work focuses on syntactic structure of an entire question to find patterns in order to classify questions. The patterns and rules are developed after using POS tagging and regular expressions. Though the work shows some reasonable findings, however there is a need to have a lot of rules in order to improve in classifying questions. The results obtained for F1 value is 77% with small data set consisting of 135 questions.

Wen-Chih Chang et al. [28] have proposed to use keyword weight technique coupled with simple keywords matching to classify questions in accordance to Bloom taxonomy. No text classifier is used in the research work. The keyword weight technique is aimed at addressing overlapping Bloom keywords that can belong to more than one Bloom taxonomy level. This keywords matching

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

only work well for simple and short open ended questions.

The work by Nazlia Omar et al. [29] is similar to the work by [27] except that the former introduces a category weighting to address the overlapping verb keywords of Bloom taxonomy that can belong to more than one Bloom taxonomy level. However this method requires human intervention in determining the weight of particular verb keyword in a sentence to a category of Bloom taxonomy and it is quite subjective as different academicians or instructors may provide different weight for the same verb keyword. Besides that, the rule based model needs to learn lots of patterns before a decent accuracy can be obtained in classifying questions.

4. QUESTION CLASSIFICATION INVOLVING OTHER LANGUAGES

Question classification does not confine to English language only but also other languages such as Chinese, Persian, Spanish, Portuguese and others. One of the earlier work in question classification using Chinese language is [3]. The main aim of the work is to accurately assign labels to question based on expected answer type. A total of 4394 Chinese questions of two-layered question taxonomy, which contains 6 coarse categories and 65 fine categories were used to classify questions. The six features which have been used in the experiment are fixed–length binary feature vector foreach Chinese question: Chinese Words Segmentation, Keyword Extraction, Bag of Words, Head Phrase, Syntactic Features, and Semantic Features. The highest test accuracy obtained for the work is 84.12% using Support Vector Machine with RBF kernel function.

In contrast to the work by [3] where it uses six features for question classification, in [7] authors have proposed to focus on semantic gram and n-gram model to achieve high accuracy in question classification. Every uni-gram consists of one word. Many Chinese words have the similar semantic meaning, such as '喜爱' and '喜欢'. Authors claim that in order to reduce the amount of uni-gram and achieve high accuracy, uni-grams which have similar semantic meaning should be grouped as a semantic uni-gram. Hence, semantic thesauruses are needed. In this work, HowNet is used to calculate the uni-grams and group those similar ones. A total of 600 Chinese questions of two-layered question taxonomy, which contains 6 coarse categories and 59 fine categories were used to classify questions. Experiments have proved that the approach can perform well when classifying the questions from the open-domain in 6 coarse categories and 59 fine categories and achieve classification accuracy up to 91.00% and 83.67% respectively. The achieved accuracy is based on semantic uni-gram and bi-gram features. Using different classifiers to evaluate the approach has also been examined. It is found that SVM based on RBF kernel outperforms other classifiers such as Naïve Bayes and SNoW with a decent accuracy of 91.00% and 83.67% over course and fine categories.

In similar work to the [7], Kepei Zhang et al.[30] have proposed to use syntactic and semantic features in question classification. The extracted features are word of sentence, POS, Named Entity, and semantics. HowNet is used to calculate semantic similarity of each word in a question in order to add semantics to feature. A total of 929 Chinese questions of two-layered question taxonomy, which contains 7 coarse categories and 50 fine categories were used to classify questions. Overall, the best classification results using SVM are as follows: the precision of coarse categories is 92%, the precision of fine categories is 85%.

In an effort to improve the work by [7] and [30], Zhang Wei et al. [31] have proposed to use Hownet and dependency syntax parsing to classify Chinese questions. The proposed extracted features are interrogative word, primary sememe, which is in HowNet, of first-degree and second-degree dependent word of interrogative word, named entity and singular/plural. The authors claim that the use of primary sememe, which is in HowNet, of first-degree and second-degree dependent word of interrogative word as feature of question classification is more reasonable than to use keywords as feature of question classification. This due to a situation in which common concept can be expressed by different words and different concepts can be expressed by the same word in different context. The first degree or second degree of dependent word revolves around the concept of dependency-relationship,i.e. binary relation of word pairs in the sentence. One of them is recorded as the head word the other is recorded as the dependent word. A total of 4238 questions were selected as training data sets and 1242 questions as testing data sets. In this work, the question taxonomy contains 6 coarse classes and 27 fine classes. With SVM, results show that the accuracy for coarse class is 93.35% while for fine class is 84.45%.

All the aforementioned work focus either rule based or statistical based approaches in question

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

classification but Ling Xia et al. [32] have integrated both approaches to classify questions in Chinese Cuisine Question Answering System. There are 4 types of categories and rule based algorithm is used to classify questions. SVM based classifier will come into play when questions fail to match any rules in rule based algorithm. The extracted features that have been used are words and POS tags. Results show that machine learning method is effective in improving the overall performance in this work. Accuracy achieved by integrating both approaches is 96.22% while merely using rule based algorithm the accuracy declines to 92.45%.

Apart from classifying Chinese questions, Ali Mollaei et al. [33] have proposed to classify Persian based language questions using Conditional Random Fields (CRF) classifier. The work takes advantages on CRF classifier with extracted features such as question informer (QI) which serves as semantic feature, words in questions, question words, N-gram, POS and other features. In this work, the question taxonomy contains 5 coarse classes and 58 fine classes. The results indicate the significant role of QI in achieving better accuracy. With QI features, the accuracy for coarse class is 80.35% while for fine class is 78.71%.

In practice, syntactic analysis using parsers and semantic analysis using HowNet or WordNet are being targeted to a particular language. In short these tools or approaches are language dependent and are specific to one particular language. A more viable idea is shifting away from language dependent method to a language independent method so that the latter method can be used for any languages in the context of question classification.

Thamar et al. [34] have proposed a language independent method for question classification. The main aim of this method is that it can be applied to different languages without requiring major modifications. As the method is language independent since no complex natural language processing tools are needed, in this work, plain lexical features which are bag of words and prefixes of words that can be extracted automatically from the questions were employed. And these features are combined with attributes whose values are obtained from the Internet. These Internet based attributes are targeted to extract evidence of the possible semantic class of the question. In order to extract the attribute, a set of heuristics is used to extract from the question a word w, or set of words, that will complement the queries submitted for the search. Thereafter a search engine will be used, in this case Google, and submit queries using the word w in combination with all the possible semantic classes. If no results are returned for any of the semantic classes then eliminating words will begin from right to left until the search engine returns results for at least one of the semantic categories. The resultant numbers from the search engine are the values for the attributes used by the learning algorithm. Results show that combining Internet-based attributes with lexical features using SVM classifier can achieve accuracy as high as 88.92%. The data set used in the experiment is DISEQuA Corpus which contains 450 questions (7 categories/classes), each one formulated in four languages: Dutch, English, Italian and Spanish. The questions are classified into seven categories: Person, Organization, Measure, Date, Object, Other and Place.

In order to improve the work in [34], the same authors in [6] have proposed to modify the heuristics which is right to left and left to right in eliminating words in the process of obtaining Internet-based attributes. Authors claim that combining elimination of words from right to left and left to right achieves better accuracy compared to elimination words from right to left alone. Results show that combining Internet-based attributes with lexical features particularly prefixes of words using SVM classifier can achieve accuracy of 78.1% and 79.11% for question classification involving Portuguese language and Spanish language respectively.

5. CONCLUSION

In this paper, a comprehensive review on question classification has been done. The findings from the most of the research work on question classification use a deep syntactic and semantic analysis instead of just using surface features such as bag-of-words and n-gram methods. This is attributed to a less information in questions compared to text documents. It has been observed that semantic features extracted from a question play a significant role in achieving a decent accuracy in classifying questions in the domain of QA and IR, though it is less apparent in the educational environment. Combining syntactic and semantic features can achieve even greater accuracy and this is proven in the research on question classification involving Chinese language. Most of the time, Support Vector Machine classifier has been employed in question classification as it performs well in working with unstructured text data. However, some research work has used CRF classifier to achieve better

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

accuracy than Support Vector Machine. Certainly this review is different from other reviews in that it covers questions classifications in a comprehensive manner involving various domains. It gives an opportunity to explore further of how certain features used in a particular domain can be potentially applied in different domains.

REFERENCES

[1] D. Metzler, W. B. Croft, "Analysis of Statistical Question Classification for Fact-based Questions," *Journal of Infomation Rerieval,* vol. 8, no. 3, 2005, pp. 481-504.

[2] X. Li. and D. Roth, "Learning question classifiers," *Proc. of the 19th International Conference on Computational Linguistics,* ACM, 2002; doi: 10.3115/1072228.1072378.

[3] X. D. Lin, H. Peng, and B. Liu, "Support vector machines for text categorization in Chinese question classification," *Proc. of the 2006 IEEE/WIC/ACM International Conference on web intelligence,* IEEE Computer Society, 2006; doi: 10.1109/WI.2006.163.

[4] D. Jurafsky, J. H. Martin, *Speech and Language Processing,* Pearson Prentice Hall, 2009.

[5] N. Yusof and C. J. Hui, "Determination of Bloom's cognitive level of question items using artificial neural network," *Proc. of 10th International Conference on Intelligent Systems Design and Applications (ISDA),* IEEE, 2010; doi:10.1109/ISDA.2010.5687152.

[6] T. Solorioi, M. Pérez-Coutiño, M. Montes-y-Gómez, L. Villaseñor-Pineda, and A. López-López, "Question classification in Spanish and Portuguese," *Proc. of the 6th International Conference on Computational Linguistics and Intelligent Text Processing,* ACM, 2005; doi: 10.1007/978-3-540-30586-666.

[7] L. Wang, H. Zhang, D. Wang, J. Huang, "Chinese Question Classification Based on Semantic Gram and SVM," *Proc. of International Forum on Computer Science Technology and Applications,* IEEE, 2009; doi:10.1109/IFCSTA.2009.111.

[8] D. Pinto, M. Branstein, R. Coleman, W. B. Croft, M. King, W. Li, and X. Wei, "QuASM: A System for Question Answering Using Semi-Structured Data," *Proc. of 2nd ACM/IEEE-CS Joint Conference on Digital libraries,* ACM, 2002; doi: 10.1145/544220.544228.

[9] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal, "Probabilistic question answering on the web," *Proc. of the 11th international conference on World Wide Web,* ACM, 2002; doi: 10.1145/511446.511500.

[10] W. W. Cohen, "Learning trees and rules with set-valued features," *Proc. of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference,* vol. 1, pp. 709–716, 1996, Menlo Park: AAAI Press/MIT Press.

[11] R. Khardon, D. Roth, L. G. Valiant, "Relational learning for NLP using linear threshold elements," *Proc. of 16th International Joint Conference on Artificial intelligence,* ACM, vol. 2, 1999, pp. 911–917.

[12] E. Hovy, L. Gerber, U. Hermjakob, C. Y. Lin, and D. Ravichandran, "Toward semantics-based answer pinpointing," *Proc. of the First International Conference on Human language technology research,* ACM, pp. 1-7, 2001; doi: 10.3115/1072133.1072221

[13] K. Hacioglu, W. Ward, "Question classification with support vector machines and error correcting codes," *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* ACM, vol. 2, pp. 28-30, 2003; doi: 10.3115/1073483.1073493

[14] D. Zhang, W. S. Lee, "Question classification using support vector machines," *Proc. Of 26th annual international ACM SIGIR conference on Research and development in information retrieval,* ACM, pp. 26-32, 2003; doi: 10.1145/860435.860443

[15] V. Krishnan, S. Das, and S. Chakrabarti, "Enhanced Answer Type Inference from Questions using Sequential Models," *Proc. of Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[16] X. Li. and D. Roth, "Learning question classifiers: the role of semantic information," *Journal Natural Language Engineering,* vol. 12, no. 3, 2006, pp. 229-249.

[17] R. May, AL, Building a Question Classifier for a TREC-Style Question Answering System, The Stanford Natural Language Processing Group, Final Projects 2004.

[18] T. T. Nguyen, L. M. Nguyen, and A. Shimazu, "Improving the Accuracy of Question Classification with Machine Learning," *Proc. of IEEE International*

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

Conference on Research, Innovation and Vision for the Future, 2007, IEEE, pp. 234-241, 2007; doi: 10.1109/RIVF.2007.369162

[19] Z. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, 2005.

[20] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Proc. of 11th Annual Conference on Computational Learning Theory*, pp. 92-100, 1998.

[21] S. Goldman and Y. Zhou, "Enhancing supervised learning with unlabeled data," *Proc. of the 17th International Conference on Machine Learning,* pp. 327-334, 2000.

[22] F. Li, X. Zhang, J. Yuan, and X. Zhu, "Classifying what-type questions by head noun tagging," *Proc. of the 22nd International Conference on Computational Linguistics,* ACM, vol. 1, pp. 481-488, 2008.

[23] J. Bullington, L. Endres, M. Rahman, "Open ended question classification using support vector Machines," *MAICS,* 2007.

[24] Z. Huang, M. Thint, Z. Qin, "Question classification using head words and their hypernyms," *Proc. of the Conference on Empirical Methods in Natural Language Processing,* ACM, pp. 927-936, 2008.

[25] T. Fei, W. J. Heng, K. C. Toh, and T. Qi, "Question Classification for E-learning by Artificial Neural Network", *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference,* 2003, pp. 1757-1761.

[26] A. A. Yahya and A. Osman, "Automatic classification of questions into Bloom's cognitive levels using support vector machines," *Proc. The International Arab Conference on Information Technology, Naif Arab University for Security Science (NAUSS)*, 2011, pp. 1-6.

[27] S. S. Haris, N. Omar, "A rule-based approach in Bloom's Taxonomy question classification through natural language processing," *Proc. of 7th International Conference on Computing and Convergence Technology (ICCCT), 2012*, pp. 410-414, Dec. 3-5, 2012.

[28] W. C. Chang and M. S. Chung, "Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items," *Proc. Joint Conferences on Pervasive Computing (JCPC), 2009,* 2009, pp. 727 – 734.

[29] N. Omar, S. S. Haris, S. Hassan, H. Arshad, M. Rahmat, and N. F. A. Zainal, "Automated analysis of exam questions according to bloom's taxonomy," *Proc. Universiti Kebangsaan Malaysia Teaching and Learning Congress 2011*, vol. 1, 2011; doi: 10.1016/j.sbspro.2012.09.278

[30] K. Zhang, and J. Zhao, "A Chinese Question-Answering System with Question Classification and Answer Clustering," *Proc. of Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2010,* IEEE, 2010; doi:10.1109/FSKD.2010.5569607.

[31] Z. Wei, C. Junjie and N. Yangqing, "Research on Chinese Question Classification Based on Hownet and Dependency Parsing," *Proc. of 3rd International Workshop on Intelligent Systems and Applications (ISA), 2011,* IEEE, 2011; doi:10.1109/ISA.2011.5873388.

[32] L. Xia, Z. Teng and F. Reng, "An Integrated Approach for Question Classification in Chinese Cuisine Question Answering System," *Second International Symposium on Universal Communication (ISUC '08),* IEEE, 2008; doi:10.1109/ISUC.2008.18.

[33] A. Mollaei, S. Rahati-Quchani, and A. Estaji, "Question classification in Persian language based on conditional random fields," *Proc. of 2nd International eConference Computer and Knowledge Engineering (ICCKE), 2012,* IEEE, 2012; doi:10.1109/ICCKE.2012.6395395.

[34] T. Solorioi, M. Pérez-Coutiño, M. Montes-y-Gómez, L. Villaseñor-Pineda, and A. López-López," A language independent method for question classification," *Proc. of the 20th International Conference on Computational Linguistics,* ACM, 2004; doi:10.3115/1220355.1220556.

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

Table 2: Overview of Question Classification In QA

Author	Classifier/Model	Features	Data Set	6-class	50-class	Remarks
David Pinto	Language models	NA	TREC	NA	NA	precision - 73% - 74%
Radev	Ripper	semantic/syntatic	TREC 8/9/10	NA	NA	error rate - 24% (TREC9)
Radev	Heuristics algorithm	semantic/syntatic	TREC 8/9/10	NA	NA	error rate - 8% (TREC9)
Li and Roth	Hierarchical classifier - SNoW	semantic/syntatic	TREC/UIUC	NA	78.8	accuracy - 84.20%
Li and Roth	Hierarchical classifier - SYN	semantic/syntatic	TREC	92.3	89.3	
Hacioglu et al.	SVM + ECOC	n-gram - bigram	UIUC/TREC	NA	80.2 - 82	
Zhang & Lee	Linear SVM	n-gram	TREC	87.4	79.2	
Zhang & Lee	SVM Tree Kernel	NA	TREC	90	NA	
Krishnan et al.	SVM + perfect informer	bigram/hypernym	UIUC	94.2	88	
Krishnan et al.	SVM + CRF	syntatic	UIUC	93.4	86.2	
Zhiheng et al.	Linear SVM	semantic/syntatic	UIUC	93.4	89.2	
Zhiheng et al.	Maximum Entropy	semantic/syntatic	UIUC	93.6	89	
Richard May et al.	Hierarchical classifier	bigram/parse signature	TREC	NA	80	
Jim Bullington et al.	Linear SVM	BOW	1000 questions	NA	NA	accuracy - 74.6%
Tri Thanh Nguyen et al.	Hierarchical classifier - MEM	BOW	TREC	NA	NA	precision - 80.4%
Fangtao Li et al.	LDCRFs based with total features	syntatic/semantic/transition	UIUC	NA	NA	accuracy - 85.60

*NA = Not Applicable/Available

Table 3: Overview Of Question Classification In Educational Environment

Author	Classifier/Model	Features	Data Set	Remarks
Ting Fei et al.	Neural Network	query-text relevance, mean term frequency, length of Q&A, term frequency and distribution of Q & A in text	233 questions	Results: F1 - 78%
Anwar Ali Yahya et al	SVM	BOW	272 questions	Results: Accuracy - 87.4%
Norazah Yusof et al.	Neural Network	whole feature set, DF feature set and CF-DF feature set	274 questions	Results: Precision - whole feature - 65.9%; DF - 62.2; CF-DF - 61.00
Syahidah Sufi et al.	Rule based classifier	syntatic - POS/regular expression	135 questions	Results: F1 - 77%
Wen-Chih Chang et al.	Keyword matching	Main verb	288 questions	Results: Correct matching items - 28%

*NA = Not Applicable/Available

Table 4: Overview Of Question Classification In Other Languages

Author	Classifier/Model	Features	Data Set	Language	Remarks
Xu-Dong Lin et al	SVM - RBF Kernel	semantic/syntatic	4394 questions - 6C 65F	Chinese	accuracy - 84.12%
Liang Wang et al.	SVM - RBF Kernel	n-gram and semantic n-gram	600 questions - 6C 59F	Chinese	accurcy - 91.00% for C; 83.67% for F
Kepei Zhang et al.	SVM	semantic/syntatic	929 questions - 7C 50F	Chinese	precision - 92% for C, 85% for F
Zhang Wei et al	SVM	HowNet/syntatic - dependency syntax parsing	4238 questions - 6C 27F	Chinese	accurcy - 93.35% for C; 84.45% for F
Ling Xia et al	Rule based/SVM	BOW/POS	NA	Chinese	accuracy - 96.22%
Ali Mollaei et al	CRF classifier	semantic/syntatic	5000 questions - 5C 58F	Persian	accurcy - 80.35% for C; 78.71% for F
Thamar et al	Linear SVM	lexical features + internet attributes	DISEQuA Corpus - 450 questions; 7 categories	Italian	accuracy - 88.92%
Thamar et al	Linear SVM	lexical features + internet attributes	DISEQuA Corpus - 450 questions; 7 categories	Portugese/Spanish	accuracy - 78.1% for Portugese; 79.11% for Spanish

*NA = Not Applicable/Available; *C - coarse categories/classes; *F - fine categories/classes