2017 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017

Question Classification Using Support Vector Machine with Hybrid Feature Extraction Method

Syed Mehedi Hasan Nirob
Computer Science and Engineering
Shahjalal University of
Science and Technology
Sylhet-3114, Bangladesh
smh.nirob@gmail.com

Md. Kazi Nayeem
Computer Science and Engineering
Shahjalal University of
Science and Technology
Sylhet-3114, Bangladesh
masum.nayeem@gmail.com

Md. Saiful Islam
Computer Science and Engineering
Shahjalal University of
Science and Technology
Sylhet-3114, Bangladesh
saiful-cse@sust.edu

***Abstract*—This paper presents an approach to categorizing Bangla language question into some predefined coarse-grained category that represents expected answer type of that particular question. Support vector machine was used with different kernel function to increase the accuracy of existing Bangla question classification system. Both predefined feature set and the stream of unigram based on the frequency of data set was considered to build feature matrix. For five cross validation average 89.14% accuracy was achieved using 380 top frequent words as the feature which outperformed existing single model based Bangla question classification system. For same cross validation, 88.62% accuracy was achieved with a combination of wh-word, wh-word position and question length as feature set.**

***Index Terms*—Question Classification, SVM, Question Taxonomy, Feature Extraction, Kernel Function, Wh-word.**

I. INTRODUCTION

Question classification is actually the system of classifying questions into some predefined class which reflects expected answer type of these questions. These semantic answer categories can also suggest different question processing strategies.

For example, the question ”Who wrote the national anthem of Bangladesh?” asks for a person name and task of a classification system is to tag this question as the person. If we find a sentence that has the answer to this question then name entity recognition of that sentence can reveal the exact answer to this question. That’s why question classification is important.

Question classification is an influential part of a question answering system [1]. A question answering system finds the most relevant answer to a question asked by a user from a lot of documents. This task is challenging because these questions are asked in natural language and don’t follow grammar rules in many cases [2]. And with a large amount of data, search space for question answering is also huge. But knowing the expected answer type can help us to reduce the search space by a considerable amount [3].

Text Retrieval Conferences question answering track has introduced different QA model with varying performance. These models use different QA framework with some form of question classification module.

II. RELATED WORKS ON QUESTION CLASSIFICATION

There are a lot of existing and ongoing research works on question classification in the different language. Research on some topic related to question classification like question classifiers, question taxonomies, question features has been issued continuously. Question feature extraction procedure and classifier used to classify question makes difference among those approach.

Rule based techniques to classify question can be less complex if we can represent question in a different way like a semantic parse tree. Authors Hermjakob et al., 2001 wrote 276 hand written rules to classify question into 122 categories [4]. But statistical question classification methods require little or no hand tuning in many instances [5]. An experiment result showed that with only surface text features like bag-of words and bag-of-n-grams the support vector machine outperforms other machine learning methods [6]. Authors Chen et al., 2006 showed that syntactic structure of a sentence can provide more convenient information than a bag of n grams [7].

Selecting an optimal set of feature has always been a challenging task for the researcher [8]. Some preferred rich feature space for their question classifier. The small-scale feature set can also be impactful if it is chosen wisely and head words are one of them [9] [10]. But authors achieved 89.2% and 89.0% accuracy using linear SVM and Maximum Entropy models with a traditional standard feature set like unigrams. On the other hand, many researchers used only n-gram as a feature with suitable rule based question classifier [11]. Authors achieved 88.8% accuracy for coarse grained categories and 80.6% accuracy for fine grained categories. Despite challenges of processing Bangla questions, there is

978-1-5386-1150-0/17/$31.00 ©2017 IEEE

some research work on Bangla question classification. In the early stage of question classification for Bangla language, only single-layer taxonomy was proposed [12]. Author's used different lexical, syntactic and semantic features and various machine learning approach to categorize nine course-grained classes. Those classifiers are Naive Bayes, Kernel Naive Bayes, Rule Induction and Decision Tree. Decision tree classifier provided highest 87.63% accuracy among all of them. Later sixty-nine fine-grained question classes for previous nine course-grained classes was suggested [13]. Machine learning ensemble technique like bagging and boosting was applied to the training data to improve accuracy for an increased number of class [14].

Research works in question classification also vary with language. We worked on Bangla question classification and result or performance won't be same even if we use a system designed for another language that uses similar feature set and algorithm.

III. DATASET PREPARATION AND ANALYSIS

It was mentioned earlier that there is no accessible Question Classification dataset for Bengali language right now. So, we collected some sample question from a website [15]. There is an existing research work that uses this dataset but their dataset is not open. This website has factoid questions in Bangla language in different categories like Bangladesh, international, literature etc.

1375 questions from Bangladesh subject and 1118 questions from international category was collected. Then we prepared 120 Bengali questions manually related to computer science. We used some selected Wikipedia article for this purpose. Both question and answer was prepared for our future analysis. From this question set we selected 1160 questions for our classification purpose. The problem with other 215 questions is that they don't represent any of the 9 category that we defined. These 1160 questions were classified into nine main categories manually. We only considered coarse-grained classes for classification. Table I shows question category details.

Table I: Bangla Question Categories

Class Name	Description
PER	Person name
LOC	Location or place related question
TIME	Time related question
GRO	Question about a group or organization
REA	Reason of something
NUM	Answer of those question will be a number
DEF	Asks for definition of something
METH	Procedure related question
MISC	Miscellaneous questions like biggest, smallest etc.

In a question dataset not every word is useful. Some data segment can make the data model unstable. Suppose, There are some English word and some special character within this dataset. We need to exclude those to improve performance.

Now, We have a question set Q with n question and for our dataset, $n = 1160$.

$$Q = \{Q_1, Q_2, Q_3 \dots Q_{n-1}, Q_n\}$$

And a set of class or category C. For our dataset, $m = 5$.

$$C = \{C_1, C_2, C_3 \dots C_{m-1}, C_m\}$$

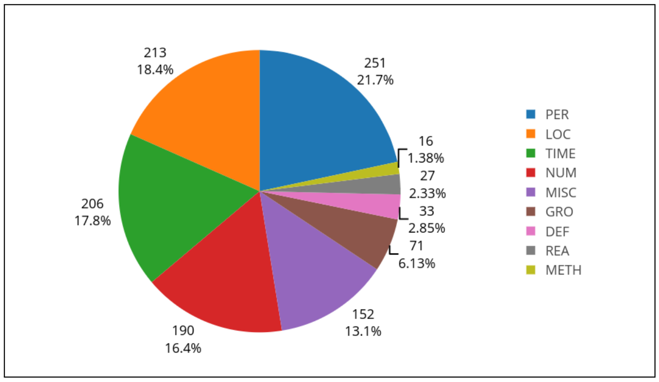

Figure 1: Number and percentage of questions in different categories

Figure 1 shows number and percentage of manually classified questions in each question category in our dataset. Number of question asking for a person name is relatively higher(21.7%) than other type of question. Then comes location type question which is 18.4% of our total dataset. Percentage of questions asking for a method of something is lowest, only 1.38% of our dataset. This type of difference in percentage has huge impact on any classification system.

IV. FEATURE EXTRACTION

Selecting an optimal set of feature is the most influential segment for any machine learning based classification model. There are several research works on feature extraction from text for categorization purpose [16].

Lets define a question Q_k with p words.

$$Q_k = W_1 W_2 W_3 \dots W_{p-1} W_p$$

W_k is any word where $1 <= k <= n$ and we selected feature based on those words.

For sentence or document level classification there are three types of feature that is needed to be consider. And these are lexical features, syntactical features and semantic features [17].

A. Lexical Features

We selected lexical features for our classifier based on words of question dataset.

Table II: WH-words in Bangla question dataset

কে	কোথায়	কিভাবে
কি	কবে	কয়টি
কিরূপ	কতটি	কারা
কত	কোন	কেন
কাকে	কোনটি	কখন
কার	কাদের	

1) Wh-word: A wh-word is one of the function word that is used to commence a wh question and a very important feature in question classification system. Sometimes only wh-word can distinguish a question category from another. If we find wh-word ”where” or "কোথায়" in a question then assuredly it can be said that this question asks for a place. So, this question will be in location category. Table refwh show’s wh-words that we extracted from out Bangla question dataset.

Although there are three types of interrogatives in Bangla language, only simple or unit interrogatives were utilized. Other two types of interrogatives are actually fusion of unit interrogatives. Because presence of dual and compound interrogatives is quite infrequent in question dataset. Hence, Compound interrogatives are irrelevant in our classification system. Suppose,

কে করে এই কাজটি করেছিল?

With interrogative কে কবে this question ask for both person and time. But according to our system one question belongs to only one class. We will choose any one of them in our system, not both of them. So, considering only unit interrogatives will work in this case.

Table III: Feature words related to question categories

Category	Words related to category
LOC	স্থান, স্থানের, দেশ, জায়গা, অবস্থিত, থানা, জেলা, দেশটিকে, দেশটি
TIME	সময়, বছর, মাস, দিন, কাল, খিষ্ট্রাব্দে, সালে, হয়েছিল
ORG	মন্ত্রনালয়, কোম্পানী, সংস্থা,কমিশন
PER	নাম, করেন, ছিলেন
NUM	সংখ্যা, পরিমান, অংশ, শতকরা, ভাগ, দূরত্ব, কততম, গড়, অবস্থান, উচ্চতা
REA	কারন, উদ্দেশ্য
MISC	পাখি, প্রাণী, বৃহত্তম, সর্বোচ্চ, দীর্ঘতম, জাতীয়, হয়, প্রথম, করে

2) Wh-word position: Wh-word position is an effective feature with wh-word. We considered four cases regarding wh-word position in question sentence.

- First position
- Second position
- Penultimate (Second to the last position)
- Last position

We noticed that in most cases position of a particular wh-word doesn’t change.

3) question length: For some particular question class length can be a critical feature. By length we mean how many word this question contains. For example, usually length of definition type question is two and number of three length location type question frequent in dataset.

B. Syntactical Features

1) Main words: In a particular question dataset, every question word is not equally important. Some word has high impact on classification system. That’s why we manually picked some word closely associated with question categories. These words occur frequently in dataset and system provides higher accuracy if used as feature. Table III shows main feature words related to question classes that was defined earlier.

There is another syntactical feature called Part of Speech(POS) tags. But we didn’t use this feature because accuracy of Bangla POS tagger is not decent. We didn’t use any semantic feature like named entities(NE) for same reason. Most importantly accuracy of our system is not dependent on any other system.

C. Other Features

Besides, training system with well defined lexical, syntactical and semantic features the first thing we tried as feature is n-gram which is actually traditional and straightforward. Individual word of a question can be very important feature space for any question classifier [18]. But if we go further with n-gram performance of classifier decreases rapidly. Bigram or trigram is not much useful to distinguish a question from other. There is another problem with unigram feature. If the number of features is much greater than the number of samples, SVM method is likely to give poor performance.

But, unigrams with higher frequency in our dataset did the trick. Higher frequency means higher impact on dataset and in question dataset we don’t have to worry about stop words. Accuracy of our system shows the proof of this observation.

Table IV: Top 10 feature words based on frequency

Feature Word	Frequency
কে	187
কোথায়	170
কোন	168
হয়	168
বাংলাদেশের	167
করেন	147
কত	124
কবে	119
কি	105
প্রথম	98

Table IV shows 10 feature words with highest frequency and of them is wh-word. If we observe frequent words list from table IV, we will find that this list has 6 wh-words from table II. Which is predictable because in a question dataset wh-words are more frequent than other words and also a vital feature candidate in question classification system.

V. METHODOLOGY

We designed our question classification system in four main steps. These are,

- Question dataset collection and processing
- Extracting feature set and building feature matrix
- Designing a machine learning based classifier
- Performance measurement

The task of question classification can be performed in two different ways. First one is hand crafted rules and the second one is using machine learning techniques. Machine learning technique was used in our research.

We have a set of question Q and a set of class or category C and our classification task is to tag questions from set Q with any one class label from set C.

After preparing question dataset, we defined optimal feature set. Then we constructed feature matrix for each feature set. In feature matrix, each row represents a question or observation and each column represents a feature. Maximum feature including n-grams is boolean in our system. Let, MAT is a feature matrix and for $MAT[i][j]$, if j'th feature is present in the i'th question then value of $MAT[i][j]$ will be 1 otherwise value will be 0. That's the main concept of our systems feature matrix.

To build a machine learning based classifier we need feature matrix and a suitable algorithm. There is no best algorithm in machine learning. Performance of an algorithm depends on the specific problem, data size, and feature set. But when it comes to text classification problems, historically performance of SVM is very decent [19] [20]. Also, besides linear classification, SVMs can efficiently map input into high-dimensional feature spaces which is called kernel trick. That's why we applied SVM algorithm in our classifier with kernel trick [21]. Given a training set of N data points $\{y_k, x_k\}_{k=1}^{N}$, where x_k is the kth input pattern and y_k is the kth output pattern, the form of classifier following support vector method approach is like Eq. 1.

$$y(x) = sign\left[\sum_{k=1}^{n} \alpha_k y_k \psi(x, x_k) + b\right] \tag{1}$$

In this equation, for every k, α_k are positive real constants and b is a real constant and $\psi(.,.)$ is the kernel function. For linear kernel function based classification system value of $\psi(x, x_k)$ is in Eq. 2.

$$\psi(x, x_k) = x_k^T x \tag{2}$$

We also tried nonlinear classification using RBF, polynomial and sigmoid kernel function. RBF or radial basis function kernel transform a single vector to a vector of higher dimensionality using Eq. 2.

$$\psi(x, x_k) = \exp(-\gamma|x - x_k||^2) \tag{3}$$

Here, x represents training question data vector and x_k is input to be classified. And γ is the slope between them.

RBF kernel is more popular in SVM classification than the polynomial kernel. But the polynomial kernel is quite popular in natural language processing or NLP than RBF. On the other hand, the polynomial kernel deals with features in a different way. To determine similarity, it looks not only at the given features of input samples, but also combinations of these features. That can improve classification performance a lot.

To evaluate a classification system we need some performance measurement technique. We measured accuracy of our system for a particular parameter set which is a widely used metrics if we need to resolve a classifiers class discrimination ability.

$$accuracy = \frac{TP + TN}{P + N}$$

Where,

TP(True Positive) = Number of positive samples and labeled as such.

TN(True Negative) = Number of negative samples and labeled as such.

P + N = Total number of positive and negative samples

We trained and evaluated our system for every combination of our feature set and kernel function. Then accuracy was measured for each parameter to find the best features-kernel combination for our classification system.

VI. RESULT AND PERFORMANCE ANALYSIS

Questions were manually classified in the predefined category for training purpose. 70% question from our dataset was used to train the system and 30% was used to measure the accuracy of the model. But to estimate a final predictive model single round of cross-validation is not enough. That's why we performed multiple rounds of cross-validation using different partitions and then the validation results were combined(averaged).

We used support vector machine with different kernel tricks to prepare the system. As feature set, wh-words are most important without any doubt. But only wh-word can't guarantee competent accuracy in most cases. We proposed some supplementary features in the previous section. Also, there are no guarantees for one kernel to work better than the other. So, we need to check every possible option and choose the best option. It is recommended to use linear kernel for text classification in some case. Because most of the text classification are linearly separable. Also, linear kernel is good when there is a lot of features. That's because mapping the data to a higher dimensional space does not really improve the performance of classification.

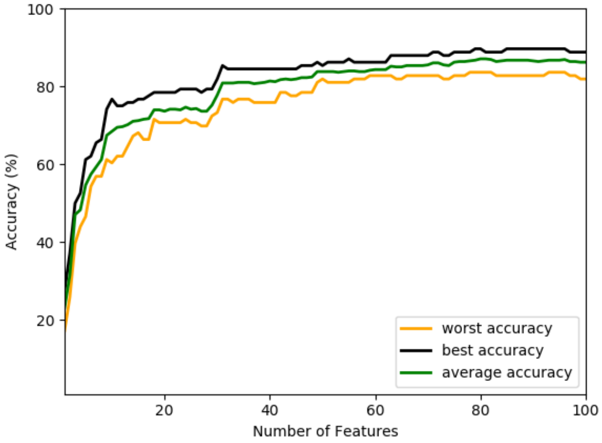

Figure 2: Performance(accuracy) for linear kernel function

At first, we experimented SVM with the linear kernel using frequent 1-grams in decreasing order. The graph in figure 2 shows the performance of SVM using linear kernel for a different number of feature. At first, accuracy is very low for less number of feature. Then, accuracy increases as the number of feature increases and at one point it becomes flat like an exponential hyperbolic curve. For 1156 most frequent unigram feature we get the best average accuracy of 89.14%. For a particular cross validation, 91.38% is the best accuracy which used 380 unigrams as a feature.

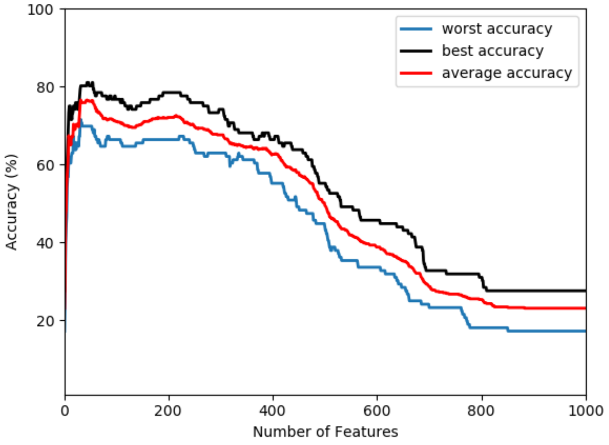

Figure 3: Performance(accuracy) for RBF kernel function

If it is not possible to separate data linearly, then we can nonlinear kernel like RBF, polynomial or sigmoid function. RBF uses normal curves around the data points and sums these so that the decision boundary can be defined for a particular class.

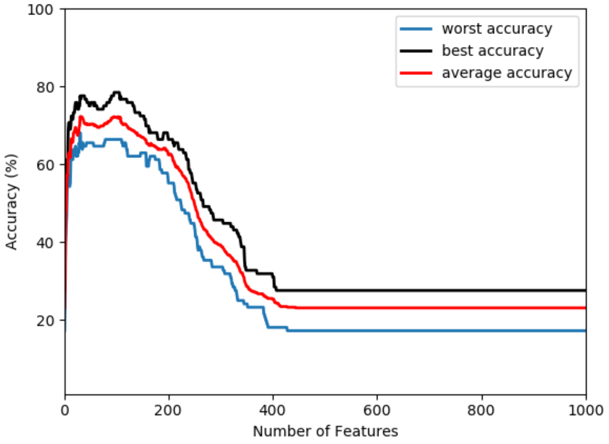

Figure 4: Performance(accuracy) for sigmoid kernel function

Figure 3 shows performance or accuracy of our system while using RBF kernel for 1 to 1000 number of features. Accuracy is much less than linear function and it decreases exponentially as a number of feature increases. Best average accuracy 76.55% was achieved for 32 most frequent words.

Figure 4 shows performance or accuracy of our system while using sigmoid kernel function for 1 to 1000 number of features. Accuracy drops more quickly than RBF kernel as a number of feature increases. Best average accuracy 72.24% was achieved for 31 most frequent words. Performance of polynomial kernel is worse than all kernel function. In best case, 41.4% accuracy can be achieved using polynomial kernel function.

After using frequent words as a feature we trained our system with the pre-defined feature set. In this time we only considered linear kernel function. In the first run, only wh-words were used as feature set. For five cross validation, 88.62% is average accuracy and 91.37% is the best accuracy. Although in the worst case of cross validation accuracy is 86.2% but average accuracy is the main fact. So, for five different test-train dataset accuracy of this system is in between 86.2% to 91.37%. But we can improve this performance by adding more feature to the feature set.

Table V: SVM linear kernel performance for 5 cross validation using specific feature set

Feature Set	Accuracy	Average Accuracy
Wh Word	86.20%	88.62%
	87.94%	
	87.94%	
	89.65%	
	91.37%	
Wh Word + Wh Word Position + Question Length	87.07%	89.31%
	87.94%	
	89.66%	
	89.66%	
	92.25%	

Later two more feature wh-word position and question length were combined with wh-word and new feature matrix was built for the classifier. And it improved the performance of our classifier by a significant amount. We achieved 89.31% average accuracy for five different cross validation or train-test data partition. In best case of dataset partition our classification system achieves 92.25% accuracy and in worst case, this accuracy decreases to 87.07% which is quite good if compared to the accuracy of existing Bangla language question classifier. Table V shows the performance of the system for this feature set.

Comparison of worst, average and best performance for the different kernel is shown on figure 5. This graph is based on frequent word feature set. From this graph, we can see that linear kernel shows better performance than nonlinear kernels like RBF, sigmoid or polynomial. Small size dataset and large size feature set is the main reason of nonlinear kernels poor performance.

Our system's average case accuracy outperformed accuracy of question classification system which relies on the single classifier. Previous best accuracy was 87.63% using decision tree classifier. And best case cross validation accuracy outperformed accuracy of ensemble approach of question classification. Four main classifiers Naive Bayes, Kernel Naive Bayes, Rule Induction and Decision Tree performance were used. But for the first time, we applied SVM or support vector machine algorithm to classify Bangla question. SVM always tries to separate different question category with most

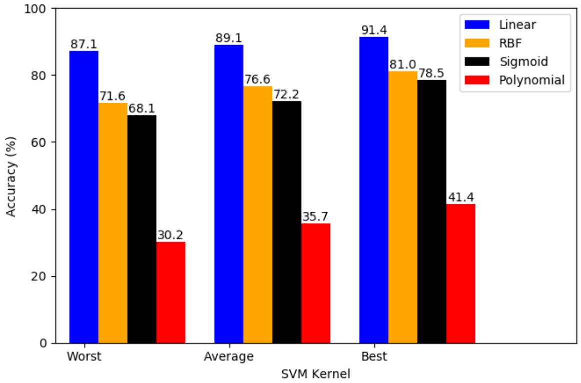

Figure 5: Worst, average and best performance comparison for different kernel

optimal hyperplane that's why it is more suitable for text categorization. Performance of our system is the proof of this assumption. Our classification is for coarse-grained or single layered classes but can be extended to fine-grained classes with ensemble approach.

Performance of a question classification system largely depends on dataset and algorithm. Our labeled small dataset can do that with help of semi-supervised learning [22]. Also the more question class we have there will be more chance of misclassification of a question. Our next target is to build a corpus with more question available. Surely that will improve current performance a lot.

VII. CONCLUSION

Our research work is a combination of support vector machine kernel function and word based feature set. Question dataset for Bangla language is not so rich to perform machine learning classification task yet we achieved the highest accuracy for an effective feature set with an efficient algorithm. It is possible to increase this accuracy with a larger dataset and applying neural network algorithm. Also, some feature like part of speech and name entity can be used for this classification task but the poor performance of Bangla language processing tools is the main problem in this case.

Question classification is a subproblem of many other problems like question answering as it represents expected answer type. Question classification can open door for many other research work in Bangla language or natural language processing field. This classification system proposes a dynamic feature selection method that can adapt to any dataset and will show better performance for a better dataset with some optimization technique.

References

[1] S. Xu, G. Cheng, and F. Kong, "Research on question classification for automatic question answering," in *Asian Language Processing (IALP), 2016 International Conference on*. IEEE, 2016, pp. 218–221.

[2] K. Yu, Q. Liu, Y. Zheng, T. Zhao, and D. Zheng, "History question classification and representation for chinese gaokao," in *Asian Language Processing (IALP), 2016 International Conference on*. IEEE, 2016, pp. 129–132.

[3] E. Haihong, Y. Hu, M. Song, Z. Ou, and X. Wang, "Research and implementation of question classification model in q&a system," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2017, pp. 372–384.

[4] U. Hermjakob, "Parsing and question classification for question answering," in *Proceedings of the workshop on Open-domain question answering-Volume 12*. Association for Computational Linguistics, 2001, pp. 1–6.

[5] D. Metzler and W. B. Croft, "Analysis of statistical question classification for fact-based questions," *Information Retrieval*, vol. 8, no. 3, pp. 481–504, 2005.

[6] D. Zhang and W. S. Lee, "Question classification using support vector machines," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 26–32.

[7] Y. Chen, M. Zhou, and S. Wang, "Reranking answers for definitional qa using language modeling," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 1081–1088.

[8] A. Sangodiah, R. Ahmad, and W. F. W. Ahmad, "A review in feature extraction approach in question classification using support vector machine," in *Control System, Computing and Engineering (ICCSCE), 2014 IEEE International Conference on*. IEEE, 2014, pp. 536–541.

[9] Z. Huang, M. Thint, and Z. Qin, "Question classification using head words and their hypernyms," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 927–936.

[10] M. Pota, M. Esposito, and G. De Pietro, "A forward-selection algorithm for svm-based question classification in cognitive systems," in *Intelligent Interactive Multimedia Systems and Services 2016*. Springer, 2016, pp. 587–598.

[11] J. Silva, L. Coheur, A. C. Mendes, and A. Wichert, "From symbolic to sub-symbolic information in question classification," *Artificial Intelligence Review*, vol. 35, no. 2, pp. 137–154, 2011.

[12] S. Banerjee and S. Bandyopadhyay, "Bengali question classification: Towards developing qa system," in *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), COLING, India*, 2012, pp. 25–40.

[13] ——, "Ensemble approach for fine-grained question classification in bengali," in *Proceedings of 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC), Taiwan*, 2013, pp. 75–84.

[14] ——, "An empirical study of combining multiple models in bengali question classification," in *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP), Japan*, 2013, pp. 892–896.

[15] "Bcs/other exam preparation," [Accessed : 11-March-2017]. [Online]. Available: http://www.bcstest.com/

[16] A. Moh'd A Mesleh, "Chi square feature extraction based svms arabic language text categorization system," *Journal of Computer Science*, vol. 3, no. 6, pp. 430–435, 2007.

[17] B. Loni, "A survey of state-of-the-art methods on question classification," 2011.

[18] M. A. Islam, M. F. Kabir, K. Abdullah-Al-Mamun, and M. N. Huda, "Word/phrase based answer type classification for bengali question answering system," in *Informatics, Electronics and Vision (ICIEV), 2016 5th International Conference on*. IEEE, 2016, pp. 445–448.

[19] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[20] S. Zadrożny, J. Kacprzyk, and M. Gajewski, "A new approach to the multiaspect text categorization by using the support vector machines," in *Challenging problems and solutions in intelligent systems*. Springer, 2016, pp. 261–277.

[21] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[22] Y. Li, L. Su, J. Chen, and L. Yuan, "Semi-supervised learning for question classification in cqa," *Natural Computing*, pp. 1–11, 2016.