An Empirical Study of Combining Multiple Models in Bengali Question Classification

Somnath Banerjee
Department of Computer Science and Engineering
Jadavpur University, India
s.banerjee1980@gmail.com

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University, India
sivaji_cse_ju@yahoo.com

Abstract

This paper demonstrates that combination of multiple models achieves better classification performance than that obtained by existing individual models for question classification task in Bengali. We have exploited state of the art multiple model combination techniques, i.e., ensemble, stacking and voting on lexical, syntactical and semantic features of Bengali question for the question classification task. Bagging and boosting have been experimented as ensemble techniques. Naïve Bayes, kernel Naïve Bayes, Rule Induction and Decision Tree classifiers have been used as base learners. The experimental results show that classifier combination models outperform existing single model approaches. Overall voting approach has achieved maximum classification accuracy of 91.65% and outperformed the existing single model approaches (maximum accuracy of 87.63%).

1 Introduction

Although different types of question answering systems (QA) have different architectures, most of them follow a framework in which question classification (QC) plays an important role (Voorhees, 2001) and QC has significant influence on the overall performance of a QA system (Ittycheriah et al., 2001; Hovy et al., 2001; Moldovan et al., 2003). The task of a question classifier is to assign one or more class labels, depending on classification strategy, to a given question written in natural language.

Basically there are two main motivations for question classification: locating the answer and choosing the search strategy. Knowing the question class not only reduces the search space needed to find the answer, it can also help to find the true answer in a given set of candidate answers.

One of the main issues of classification modeling is the improvement of classification accuracy. For that purpose, many researchers have recently placed considerable attention to the task of classifier combination methods. The idea is not to rely on a single decision making scheme. Instead, many single classifiers are used for decision making by combining their individual opinions to arrive at a consensus decision.

2 Related Work and Motivations

A lot of researches on QC, question taxonomies, and question features are being published continuously. There are basically two different approaches used to classify questions- one is rule based (Hull, 1999; Prager et al., 1999) and another is machine learning based (Zhang et al., 2003; Li and Roth, 2004). However, a number of researchers have also used some hybrid approaches which combine rule-based and machine learning based approaches (Huang et al., 2008; Silva et al., 2011).

Many researchers have investigated the technique of combining the predictions of multiple classifiers to produce a single classifier (Breiman, 1996; Clemen, 1989; Perrone, 1993; Wolpert, 1992). The resulting classifier is generally more accurate than any of the individual classifiers making up the ensemble. Both theoretical (Hansen and Salamon, 1990; Krogh and Vedelsby , 1995) and empirical (Hashem, 1997; Opitz and Shavlik, 1996a, 1996b) researches have been carried out successfully. Last decade a group of researchers focused on classifier combination methods in question classification task. LI *et al.* (2005) trained four SVM classifiers based on four different types of features and combined them with various strategies. Later LI *et al.* (2006) performed similar type of experiments and achieved

International Joint Conference on Natural Language Processing, pages 892–896,
Nagoya, Japan, 14-18 October 2013.

improved accuracy on TREC dataset. (Jia et al., 2007; Su et al., 2009) proposed ensemble learning for Chinese question classification.

Recently, (Banerjee and Bandyopadhyay, 2012) have worked on Bengali QC task and achieved 87.63% accuracy using single classifier approach. So far, classifier combination methods have not been used by any researcher in Bengali question classification task. So, we employ the use of classifier combination methods to improve question classification accuracy.

3 Question Type Taxonomies

The present work follows the QC taxonomies proposed by (Banerjee and Bandyopadhyay, 2012) for two reasons. First, that is the only standard taxonomy that exists in Bengali QC so far. Secondly, the results of the present work can be compared with the work of (Banerjee and Bandyopadhyay, 2012) to establish the improvement in accuracy.

4 Features

In the task of question classification, there is always an important problem to decide the optimal set of features to train the classifiers. Different studies have extracted various features with different approaches. The features in question classification task can be categorized into three different types: lexical, syntactical and semantic features (Loni, 2011).

Loni *et al.* (2011) also represented a question in the QC task similar to document representation in vector space model, i.e., a question is a vector which is described by the words inside it. Therefore a question Q can be represented as:

$$Q = (W_1, W_2, W_3, ..., W_{N-1}, W_N)$$

Where, W_K= frequency of term K in question Q, and N= total number of Terms

We have also used three types of features for QC. We use the same features previously used by (Banerjee and Bandyopadhyay, 2012).

Lexical features (f_{Lex}): wh-word, wh-word positions, wh-type, question length, end marker, word shape.

Syntactical features (f_{Syn}): POS tags, head word.

Semantic features (f_{Sem}): related words, named entity.

5 Combined Model Learning for QC

There are three approaches of classifier combination: 1) Ensemble, 2) Stacking and 3) Voting.

Two popular methods for creating accurate ensembles are *bagging* (Breiman, 1996) and *boosting* (Freund and Schapire, 1996; Schapire, 1990). We have used *Rapid Miner*[1] tool in the experiments of this work.

6 Experiments

This section describes our empirical study of *ensemble*, *stacking* and *voting* approaches. Each of these three approaches has been tested with Naïve Bayes (NB), Kernel Naïve Bayes (k-NB), Rule Induction (RI) and Decision Tree (DT). The previous work (Banerjee and Bandyopadhyay, 2012) on Bengali question classification task used these four classifiers. So in this work, we have used those classifiers to establish the effect of combining models.

6.1 Dataset

The present research work adopts the same corpus used by (Banerjee and Bandyopadhyay, 2012). The corpus consists of 1100 Bengali questions of different domains, e.g., education, geography, history, science etc. We have used 770 questions (70%) for training and rest 330 questions (30%) to test the classification models.

6.2 Results

In total thirteen different experiments have been performed. Four different experiments have been performed for each *bagging* and *boosting*. So, altogether eight different experiments have been performed for the ensemble approach. Four different experiments have been performed for *stacking*. But for *voting*, a single experiment has been performed. Actually, each experiment can be thought of as three experiments, because a classifier model has been tested on f_{Lex}, $f_{Syn} + f_{Sem}$ and $f_{Lex} + f_{Syn} + f_{Sem}$ features separately. The outcome of the experiments have been tabulated and described in the next sub-sections.

In our study, *classification accuracy* has been used to evaluate the results of the experiments. *accuracy* is the widely used evaluation metric to determine the class discrimination ability of classifiers, and is calculated using the following equation:

$$accuracy(\%) = \frac{T_P+T_N}{P+N}$$

[1] http://www.rapidminer.com

where, T_P = true positive samples; T_N = true negative samples; P = positive samples; N = negative samples.

It is a primary metric in evaluating classifier performances and it is defined as the percentage of test samples that are correctly classified by the algorithm.

6.2.1 Results based on Bagging

Bagging approach has been applied separately to four classifiers (i.e., NB, k-NB, RI and DT) and Table-1 tabulates the detailed information of the accuracy obtained.

BL	f_{Lex}	f_{Lex}+f_{Syn}	f_{Lex}+f_{Syn}+f_{Sem}
NB	81.53%	82.77%	83.25%
k-NB	82.09%	83.37%	84.22%
RI	83.96%	85.61%	86.90%
DT	85.23%	86.41%	**91.27%**

Table 1: Experimental results of Bagging.

Initially the size (number of iteration) of the base learner is set to 2. Then experiments have been performed with gradually increased size (size>2). The classification accuracy has been increased with increase in size. But after a certain size, the accuracy has been almost stable. At size=2 and feature=$f_{Lex} + f_{Syn} + f_{Sem}$, the NB classifier achieves 82.23% accuracy and at size$>=$ 9, it becomes stable with 83.25% accuracy. At size=2 and feature=$f_{Lex} + f_{Syn} + f_{Sem}$, the k-NB classifier achieves 83.87% accuracy and at size$>=$15, it becomes stable with 84.22% accuracy. At size=2 and feature=$f_{Lex} + f_{Syn} + f_{Sem}$, the RI classifier achieves 85.97% accuracy and at size$>=$8, it becomes stable with 86.90% accuracy. At size=2 and feature=$f_{Lex} + f_{Syn} + f_{Sem}$, the DT classifier achieves 88.09% accuracy and at size$>=$7, it becomes stable with 91.27% accuracy. It has been observed from the experiments that at each case Bagging with DT requires less size, i.e., less iteration then the other used classifiers. For experiment with f_{Lex} features, the bagging size of NB, k-NB, RI and DT are 12, 19, 11 and 10 respectively after which classification accuracy becomes stable. And For experiment with $f_{Lex} + f_{Syn}$ features, the bagging size of NB, k-NB, RI and DT are 10, 17, 9 and 8 respectively after which classification accuracy becomes stable.

6.2.2 Results based on AdaBoost.M1

Like bagging, AdaBoost.M1 has also been applied separately to the four classifiers (i.e., NB, k-NB, RI and DT). Table-2 tabulates the detailed information of the accuracy obtained.

Here, we empirically fix the iterations of AdaBoost.M1 for four classifiers to 12, 16, 10 and 8 respectively for features=$f_{Lex} + f_{Syn} + f_{Sem}$, because the weight of 1/β_t is less than 1 after those values. If 1/β_t is less than 1, then the weight of classifier model in boosting may be less than zero for that iteration.

BL	f_{Lex}	f_{Lex}+f_{Syn}	f_{Lex}+f_{Syn}+f_{Sem}
NB	81.74%	82.71%	83.51%
k-NB	83.97%	85.63%	86.87%
RI	83.55%	85.59%	86.27%
DT	85.21%	86.58%	**91.13%**

Table 2: Experimental results of AdaBoost.M1.

Similarly, for features=$f_{Lex} + f_{Syn}$ and features=f_{Lex} the iterations are 13, 18, 12, 9 and 14, 19, 14, 11 respectively for four classifiers correspondingly. The experiment results show that the performance of k-NB classifier has been improved over RI. But, overall DT performs better than all.

6.2.3 Results based on Stacking

In stacking, out of four classifiers three classifiers have been used as the *base learner* (BL) and the remaining classifier has been used as *model learner* (ML). So, four experiments have been conducted separately where each classifier get a chance to be the *model learner*. Table-3 shows the detailed information of the accuracy obtained.

BL	**ML**	f_{Lex}	f_{Lex}+f_{Syn}	f_{Lex}+f_{Syn}+f_{Sem}
k-NB,RI,DT	NB	81.76%	82.79%	83.64%
NB, RI, DT	k-NB	83.86%	85.54%	86.75%
NB,k-NB,DT	RI	85.55%	87.69%	**91.32%**
NB,k-NB,RI	DT	85.07%	86.73%	89.13%

Table 3: Experimental results of Stacking.

In the first experiment, three classifiers k-NB, RI and DT have been selected as the *base learner*s and the NB classifier has been selected as the *model learner*. Similarly, four experiments have been done selecting k-NB, RI and DT as *model learner* respectively. Experimental results show that with RI as the *model learner* and NB, k-NB, DT as the *base learner*s, the classifier achieves best classification accuracy.

6.2.4 Results Based on Voting

In voting, four classifiers altogether have been used as the *base learners* and *majority vote* has been used as voting approach. Table 4 tabulates the detailed information of the accuracy obtained.

BL	f_{Lex}	f_{Lex}+f_{Syn}	f_{Lex}+f_{Syn}+f_{Sem}
NB, RI, k-NB,DT	86.59%	88.43%	**91.65%**

Table 4: Experimental results of Voting.

7 Conclusions and Perspectives

The automated Bengali question classification system by (Banerjee and Bandyopadhyay, 2012) is based on four classifiers namely Naïve Bayes, Kernel Naïve Bayes, Rule Induction and Decision Tree. Table-5 tabulates the detailed information of the accuracy obtained.

BL	f_{Lex}	f_{Lex}+f_{Syn}	f_{Lex}+f_{Syn}+f_{Sem}
NB	80.65%	81.34%	81.89%
k-NB	81.09%	82.37%	83.21%
RI	83.31%	84.23%	85.57%
DT	84.19%	85.69%	**87.63%**

Table 5: Experimental results of (Banerjee and Bandyopadhyay, 2012)

Naïve Bayes has been used as the baseline and they have achieved 87.63% accuracy using Decision Tree. But, they have used each classifier as single model separately. The present work shows that classifier combination technique can improve the performance of question classification. Each classifier combination model performs well than single classifier model in terms of classification accuracy.

If we compare the results of previous experiment (Banerjee and Bandyopadhyay, 2012) with *bagging* approach, then classification accuracy of all the classifiers have been notably increased. The classification accuracy on f_{Lex} ,$f_{Lex} + f_{Syn}$ and $f_{Lex} + f_{Syn} + f_{Sem}$ features have been increased by 1.04%, 0.72% and 3.64%. Similarly in the *boosting* approach, the classification accuracy of all the classifiers have been notably increased and on f_{Lex} ,$f_{Lex} + f_{Syn}$ and $f_{Lex} + f_{Syn} + f_{Sem}$ features the classification accuracy have increased by 1.02%, 0.89% and 3.50%. *Stacking* approach notably increases the accuracy on $f_{Lex} + f_{Syn}$ features than *bagging* and *boosting* approaches. The classification accuracy on f_{Lex} ,$f_{Lex} + f_{Syn}$ and $f_{Lex} + f_{Syn} + f_{Sem}$ features have been increased by 1.36%, 2.74% and 0.69% respectively. *Voting* approach not only increases the classification accuracy but also hits the maximum accuracy on all features than other combined approaches. *Voting* approach increases the classification accuracy on f_{Lex} ,$f_{Lex} + f_{Syn}$ and $f_{Lex} + f_{Syn} + f_{Sem}$ features by 2.40%, 2.40% and 4.02% respectively.

So, overall *voting* approach with *majority voting* has performed best among all four classifiers combination approaches namely *bagging*, *boosting*, *stacking* and *voting*. Experimental results show that classifiers combination approaches outperform the previous single classifier classification approach by (Banerjee and Bandyopadhyay, 2012) for Bengali question classification.

The main future direction of our research is to exploit other lexical, semantic and syntactic features for Bengali question classification. In future an investigation can be performed on including new Bengali interrogatives using a large corpus. It is also worth investigating fine-grained classes for Bengali questions. In the current work, we have only investigated the Bengali questions. But, this work can be applied to other languages having low resources.

Acknowledgments

We acknowledge the support of the Department of Electronics and Information Technology (DeitY), Ministry of Communications and Information Technology (MCIT), Government of India funded project "*Development of Cross Lingual Information Access (CLIA) System Phase II*".

References

Abraham Ittycheriah , Franz Martin, Zhu Wei-Jing, Adwait Ratnaparkhi, and Richard J. Mammone. 2001. *IBMs statistical question answering system.* In Proceedings of the 9th Text Retrieval Conference, NIST.

Anders Krogh and Jesper Vedelsby. 1995. *Neural network ensembles, cross validation, and active learning.* Advances in neural information processing systems, Vol. 7, pp. 231-238 Cambridge, MA. MIT Press.

Babak Loni. 2011. *A survey of state-of-the-art methods on question classification*. Delft University of Technology, Tech. Rep (2011): 1-40.

Babak Loni, Gijs van Tulder, Pascal Wiggers, Marco Loog, and David Tax. 2011. *Question classification with weighted combination of lexical, syntactical and semantic features*. TSD, pages 243-250.

Dan Moldovan, Marius Pasca, SandaHarabagiu, and MihaiSurdeanu. 2003. *Performance issues and error analysis in an open-domain question answering system*. ACM Trans. Inf. Syst., 21:133-154.

David A. Hull. 1999. *Xerox TREC-8 question answering track report*. In Voorhees and Harman.

David H. Wolpert. 1992. *Stacked generalization*. Neural Networks, 5, 241-259.

David W. Opitz and Jude W. Shavlik. 1996a. *Actively searching for an effective neural network ensemble*. Connection Science, 8(3/4): 337-354.

David W. Opitz and Jude W. Shavlik. 1996b. *Generating accurate and diverse members of a neural network ensemble*. Advances in Neural Information Processing Systems, Vol. 8, pp. 535-541 Cambridge, MA. MIT Press.

Dell Zhang and Wee Sun Lee. 2003. *Question classification using support vector machines*. ACM SIGIR, pages 26-32,New York,USA, ACM.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin yew Lin, and Deepak Ravichandran. 2001. *Toward semantics-based answer pinpointing*.

Ellen M. Voorhees. 2001. *Overview of the TREC 2001 question answering track*. TREC, pp. 42-51.

Joao Silva, Luisa Coheur, Ana Mendes, and Andreas Wichert. 2011. *From symbolic to sub-symbolic information in question classification*. Artificial Intelligence Review, 35(2):137-154.

John Prager, Dragomir Radev, Eric Brown, and Anni Coden. 1999. *The use of predictive annotation for question answering in trec8*. TREC-8,pp.399-411.NIST.

Keliang Jia, Kang Chen, Xiaozhong Fan, Yu Zhang. 2007. *Chinese Question Classification Based on Ensemble Learning*. ACIS. pp. 342-347.

Lars Kai Hansen, and Peter Salamon. 1990. *Neural network ensembles*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 993-1001.

Lei Su, Hongzhi Liao, Zhengtao Yu, Quan Zhao. 2009. *Ensemble Learning for Question Classification*. ICIS 2009. pp. 501-505.

Leo Breiman. 1996. *Stacked regressions*. Machine Learning, 24(1), 49-64.

LI Xin, Xuan-Jing HUANG, and Li-de WU. 2006. *Question Classification by Ensemble Learning*. IJCSNS, 6(3), page : 147.

Michael Peter Perrone. 1993. *Improving Regression Estimation: Averaging Methods for Variance Reduction with Extension to General Convex Measure Optimization*. Ph.D. thesis, Brown University, Providence, RI.

Robert E. Schapire. 1990. *The strength of weak learnability*. Machine Learning, 5(2), page:197-227.

Robert T. Clemen. 1989. *Combining forecasts: A review and annotated bibliography*. International Journal of Forecasting 5, no. 4: 559-583.

Sherif Hashem. 1997. *Optimal linear combinations of neural networks*. Neural Networks, 10 (4), pp:599-614.

Somnath Banerjee and Sivaji Bandyopadhyay. 2012. *Bengali Question Classification: Towards Developing QA System*. In Proceedings of SANLP-COLING, pages 25-40, Mumbai, India.

Somnath Banerjee and Sivaji Bandyopadhyay. 2012a. *Question Classification and Answering from Procedural Text in English*. In Proceedings of QACD-COLING, pages 11-26, Mumbai, India.

Xin Li and Dan Roth. 2004. *Learning question classifiers: The role of semantic information*. COLING, pp. 556-562.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. *Question classification using head words and their hypernyms*. EMNLP, pp. 927-936.