Shiraz: A Proposed List Wise Approach to Answer Validation

Amin Heydari Alashty Saeed Rahmani Meysam Roostaee Mostafa Fakhrahmad
Shiraz University
Eram Street
Shiraz, Iran
heidari@cse.shirazu.ac.ir
{srahmani,mroostaee,mfakhrahmad}@shirazu.ac.ir

Abstract

Answer Validation is an important step in Automatic Question Answering systems and nowadays by spreading Community Question Answering systems it is known as an important task by itself. Previous works just considered it as a binary classification problem in which they try to find the best answer among all the candidate answers for a question. Accordingly, they do not consider the possible unique information which may have been included other answers. This can be considered by having a multiclass label classification problem, it is not only able to find the best answer but also can find "potentially good", "bad", and etc. answers too. By doing so, it is fully expected to extract and rate all the necessary information from existing candidates to help questioner to find the best and general answer for his question. This work tries to consider some features which are gained from importance of comments of the questioner. Finally, by using a good classifier, we try to overcome this problem. The designed system participated in subtask A of the Semeval-2015 Task 3. The primary submission ranked at the 5th and 7th places in four class label and three class label evaluation, accordingly.

1 Introduction

By spreading Community Question Answering (CQA) systems, there have been created a new taxonomy for Question Answering (QA) systems: Regular QAs, and CQAs. A regular QA, accepts a natural language input and after searching into it's available resources, returns the best shortest answer, it could find. In these systems, answering to factoid questions may be an easier challenge than the other question types. One of the features of CQA systems is its users. Once one asks a question, others try to answer that question. Then these kinds of systems just try to use users knowledge to answer users questions. Of course instead of finding the correct answer of an asked question from some candidate answers which must be done by questioner, system tries to tell the questioner which answer is helpful and which one is not. Then discussing about factoid questions is maybe so hard and it could not be handled just by using the answers and questions body, rather, it should have access to a great knowledge source to check if an answer is correct or not.

Community Question Answering systems' spreaded over the internet, and accordingly, it made researchers to be interested in getting involved to the challenges related to these systems. One of the main challenges which may be so important in the aspect of all the people who are using these systems is Answer Validation. More researches has been done as CQA systems are getting more and more popular. This challenge is a kind of classification problem which classifies comments of a question and by doing so, it can help questioner to find the correct answer, sooner, and without spending so much time to read all the comments. Alternately, it can help other web users who had searched for the similar question in a search engine and redirected to our website, to find the answer they are looking for. Next it can help us to find the questions without any proper answer, and in addition it can be used for question routing challenge (Gkotsis et al., 2014).

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 220–225,
Denver, Colorado, June 4-5, 2015.

Eventually its important to CQA systems owners to attract more users and accordingly, attracting more users means earning more money.

In this work a new type of features will be discussed which could be gained by considering the information of questioner comments. Experiments shows, this kind of features are more valuable in contrast of the most valuable features of previous works.

Some previous works focused on the deep textual features such as syntactic, lexical, and discourse features to find the best answer. And some others tried to overcome this problem using shallow features such as word count in an answer, answer count for a question, (Gkotsis et al., 2014; Toba et al., 2014). Some others, tried to propose a solution by using reputational features of such a system like user rating (a high ranked user may produce a more reliable answer), Answer rating (an answer with more ratings from other users may be more reliable), (Anderson et al., 2012).

Of course previous works, mostly have just tried to find the best answer (designed a binary classifier) but present work classifies answers into six classes: *Good*, *Potential*, *Bad*, *Dialogue*, *Not English*, and *Other*. *Good* is a comment with a complete bunch of relevant information. *Potential* is a comment with some helpful information but is not a complete answer. *Bad* is a comment with no helpful information to answer the question. *Dialogue* is a comment which shows a kind of discussion between users and obviously contains no useful information. *Not English* is a comment in other languages. *Other* is a comment which is not a kind of above mentioned classes. Samples of *English*, and *Other* classes have no valuable information as samples related to *Bad* and *Dialogue* classes.

The remainder of the paper is organized as follows: related works are presented at section 2. There is an introduction to the used dataset at section 3. At section 4 the Features will be introduced. At section 5 experiments are discussed. Finally, Section 6 would have a conclusion.

2 Related Works

In (Jeon et al., 2006) there was an attempt to overcome this challenge using non-textual features. Non-textual features are acclaimed to have lots of information which can be helpful for finding class label of an answer. Its pointed that a not properly usage of these features is the cause to not have good results. For feature selection they had estimated the correlation between the feature values and the manually judged quality scores. Higher correlation means the feature is a better indicator to predict the quality of answers. Then because Maximum Entropy models need monotonic features a feature converter was used. KDE (Kernel Density Estimation) is the one which is used in this work. At last they could get a better performance than the random ranker.

In (Shah and Pomerantz, 2010) the goal is to predict if an answer was chosen by the questioner as the best answer. They have just used features related to answers, because question's features were not that much effective. Experiments were done twice: first by estimating features' values using Amazon Turk, and second by using values automatically generated from source of questions and answers and users profiles. The results show that using second approach is more useful. First approachs features are so correlated and cannot model the variability in the data but the second approachs model is quite good in terms of its power to explain the variability in the data.

(Wang et al., 2009) proposed an analogical reasoning-based approach to measure the analogy between the new question-answer linkages and those of previous relevant knowledge which only contains positive links. And the most analogous link was assumed to be the best answer. There is an assumption that provides each answer is connected to its question with various types of latent links. Positive links indicating high-quality answers and Negative links indicating incorrect answers or user-generated spam. This work tried to solve problem of lexical gap between questions and answers. To do so, similar question and answer pairs from available questions and their correct answers in the system were utilized.

In (Surdeanu et al., 2011) linguistic features were used to represent content. The proposed method is called FMIX (feature mix), that is a mixture of four types of features: Similarity Features, Translation Features, Density/Frequency Features, and Web Correlation Features. Value of these set of features estimated using a generative model but a discrimi-

native model (SVM, Perceptron are used) was used to combine them.

In (Gkotsis et al., 2014) some shallow textual features (like answer count, longest sentence length and any other feature which does not need that much effort to retrieve from text like semantic or syntactic features) had been mainly considered. Experimental results for different mixtures of mentioned features and some other types like reputational features (e.g. answer rating, question rating) had been estimated to confirm a suitable usage of shallow features can result a prosper approach. This work's main contribution is proposing a discretization method to solve language evolution, generality problems, and accuracy. The discretization method is consists of three steps: grouping (group answers related to a question), sorting (sort answers according to their value for that feature), and discretization (assign a rank for each answer, starting from 1 and incrementing this rank by one).

In (Toba et al., 2014) a 2-layer classification method has been proposed. The first layer just tries to find the type of the question and the second layer uses the result of the first layer to find the best answer of the question. For each question type there is a specific classifier at the second layer, and furthermore a feature set which consists of a mixture of shallow and deep textual features and reputational features.

3 Dataset

The source of the corpus is the Qatar Living Forum data[1]. Details of the method of extracting and labeling its content are described at (Màrquez et al., 2015). This corpus was provided into three parts: train set, development set, and test set. And for two sub-tasks. Each of the mentioned sets is consists of a number of questions and for each question, there is some comments.

4 Features

In this section, features used for training and testing the classifier are introduced. Some shallow textual features are considered. Alternatively, we tried to extract and use reputational features as well. Some of the shallow features used, are the same as shallow features in (Gkotsis et al., 2014; Toba et al., 2014) and some other features are from the available information in the corpus like: Creation Date, Category, and Question Type. It was assumed Questioners comments can be so informative, experiments show that, features which are using this fact can be so effective.

4.1 Reputational Features

An important part of CQA systems is users reputational information. There are some previous works used the authority of the users like Anderson (2012). There is somehow no explicit information in our train set to have features of this type. But by knowing that there is an overlap between user set whose questions or comments are presented in train set and in test set two features were added to cover this type:

- **Which User Group**: gives to all comments of a certain user a unique identifier.
- **Which User Category**: gives to each comment of a certain user in each category a unique identifier.

4.2 List Wise Features

Some approaches tried to use some kinds of prior knowledge like previous available questions and their comments in system. Some others without caring about that knowledge just tried to overcome this problem using the information exists in domain of a question. In this work the most important extracted feature is presented in this type. Its according to the fact that, valuable information can be gained from differentiating questioner and commenters comments. At first we used 2 features to use this information and we were hopeful that our machine learning method can detect the relationship between these two features:

- **Questioner Id**: questioner identifier which is represented by QUSERID in dataset.
- **Commenter Id**: commenter identifier which is represented by CUSERID in dataset.

But disappointingly, those methods could not detect relationships. Then one aspect of their relationships is used and ids eliminated:

[1] http://www.qatarLiving.com/forum

- **Is Commenter Asker**: its a binary feature. *Zero* would be assigned to a comment if its CUSERID is different from QUSERID of the corresponding question. Then one would be assigned to a comment which its CUSERID is the same as the QUSERID.

Emperical results show that, this feature can seperate samples of "Dialogue" class in an acceptable rate. When a questioner make a comment, this comment can be classified into different classes as below:

- **Dialogue**: If questioner just wants to express his opinion about previous comments to his question or may be in another case, if questioner is comminicating with other users about his question using comments, and may be some other cases this comments can be classified as Dialogue class.
- **Good, Potential**: If questioner himself had found the correct answer or at least the his expected answer, he can make a comment to share the answer to other and again in this case and may be some other cases this kind of comment can be classified into Good or Potential classes.
- **Bad**: Questioner even can make a bad comment. It can has some reasons like: if he had been hopeless of receiving any response from other users then this situation can make him to post a irrelevant comment which can not help to find the answer of question.

Of course, its believed that this feature is not the true complete potentiality of the mentioned fact. There is a ranking between all the above discussed features in Table 1 according to their Gain Ratio. Answer Count is the feature with the best Information Gain (IG) in Gkotsis (2014). But it's obvious that the "Is Commenter Asker" which is a List Wise feature has gained a much better Gain Ratio from other features.

Feature	Gain Ratio
Is Commenter Asker	0.18002
Answer Count	0.0431
Type	0.03762
Category	0.01835
Length	0.01678
Avg Word Per Sentence	0.01671
Avg Char Per Sentence	0.01503
Longest Sentence	0.01296
Which User Group	0.00847
Creation Date	0.00817
Which User Category	0.0042

Table 1: General Features Gain Ration.

5 Experiments

5.1 Learning Method

Different kinds of learning methods had been tested to find the best method. At last, J48 method could result better than the others. Then it used in a bagging method. Weka (Hall et al., 2009) was used to apply learning methods to extracted features. The overall configurations in Weka are:

Bagging -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 – -C 0.25 -M 10

Before test set release time, 10-Fold cross validation was used for system evaluation. (-I 10) Experiments shown that the best minNumObj option in J48 method is 10 for this problem. (-M 10)

5.2 Discussion

As previously mentioned, CQA systems dataset are unbalanced. According to this fact, two types of train data has been generated from questions and comments. First one has the same number of comments and Second one is generated from the first set, of course with additionally redundant smaples. For each class, redundant samples have been added till its samples number get equal to the majority class. The first model was submitted as contrastive1 and the second model was the primary submission.

There was two ways of evaluation in this task. First one maps "Dialogue", "Not English", "Other" class labels to "Bad" class label, and this was called "COARSE EVALUATION" and official ranking of teams was done according to this measurement. And the second one maps just "Not English", and "Other" class labels to "Bad" class label, and it was called "FINE-GRAINED EVALUATION".

Shiraz group's primary submission has gained two different ranking according to each of the eval-

uation methods. According to fine-grained evaluation, we were ranked as the 5th, and according to coarse evaluation, were ranked as the 7th team, and the latter ranking is our official ranking for subtask A. For each of the groups two measure were estimated: F1-Score and Accuracy. Groups were ranked according to F1-Score. Shiraz's two most important submissions for each of the evaluation methods measurements are shown in Table 2.

Most of previous works had just tried to improve accuracy of their system, but using macro-F1 as the measurement of official ranking has shown that considering accuracy in this problem which has multi class labels, and data is imbalance can not be a good idea. For example, there may be a system which just tries to cover classes with majority samples in data set then it is expected to improve accuracy but it can not ensure that it could gain a suitable macro-F1. It's because that system may not be able to classify correctly samples of other classes. It means, the best system is the one which could has the best behaviour in all the classes not just some of them.

At last, it needs to be mentioned that the list wise approach is not limited to a special kind of features like textual or non-textual features. Of course, it can help to extract some new features which are so helpful to improve the classifier.

	F1-Score	Accuracy
Prm[2]_Coarse	47.34	56.83
Contr[3]_Coarse	45.03	62.55
Prm_Fine	40.06	48.53
Contr1_Fine	37.77	55.16

Table 2: System Evaluation Measure values.

The most important point in Gkotsis (2014) is discretization method. That method had been used for some continuous shallow features, but as can be seen in Table 3 F1-Score is not improved. Then the discretization method described in Gkotsis (2014) is not useful for this problem on this dataset.

[2]Primary

[3]Contrastive

	F1-Score	Accuracy
UnBalanced_Coarse	42.85	61.74
Balanced_Coarse	25.89	36.84
UnBalanced_Fine	36.61	52.83
Balanced_Fine	21.09	23.48

Table 3: System evaluation measure value for discretized Feature values.

6 Conclusion

By widely spreading of Community Question Answering systems, solving challenges of these systems is essential. The proposed system aims to improve previous solutions for Answer Validation using some new valuable features. Moreover, questioners comments have been introduced as a source of feature which can be used for extracting more powerful features from it. Only one feature was extracted using this source in this work, but it was the most valuable one. Using just a few number of features Shiraz system could gain an acceptable ranking.

As mentioned before in this kind of problems F1-score is the main measurement which should be improved in designig a system, but empirically it was shown that discretization is not helpful to achieve this goal.

Acknowledgments

We thank Living Qatar for providing data and Semeval task organizers for organizing this problem.

It is essential to thank Dr. Hooman Tahayori, Abolfazl Moridi, and all other people help us in this work with their comments.

References

Anderson, Ashton and Huttenlocher, Daniel and Kleinberg, Jon and Leskovec, Jure (2012). Discovering value from community activity on focused question answering sites: a case study of stack overflow. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 850-858).

Gkotsis, George and Stepanyan, Karen and Pedrinaci, Carlos and Domingue, John and Liakata, Maria (2014). It's all in the content: state of the art best answer prediction based on discretisation of shallow

linguistic features. In Proceedings of the 2014 ACM conference on Web science (pp. 202-210).

Hall, Mark and Frank, Eibe and Holmes, Geoffrey and Pfahringer, Bernhard and Reutemann, Peter and Witten, Ian H. (2009). The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1), (pp. 10-18).

Jeon, Jiwoon and Croft, William Bruce and Lee, Joon Ho and Park, Soyeon (2006). A framework to predict the quality of answers with non-textual features. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 228-235).

Màrquez, Lluís and Glass, James and Magdy, Walid and Moschitti, Alessandro and Nakov, Preslav and Randeree, Bilal (2015). SemEval-2015 Task 3: Answer Selection in Community Question Answering. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).

Shah, Chirag and Pomerantz, Jefferey (2010). Evaluating and predicting answer quality in community QA. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 411-418).

Surdeanu, Mihai and Ciaramita, Massimiliano and Zaragoza, Hugo (2011). Learning to rank answers to non-factoid questions from web collections. Computational Linguistics, 37(2), (pp. 351-383).

Toba, Hapnes and Ming, Zhao-Yan and Adriani, Mirna and Chua, Tat-Seng (2014). Discovering high quality answers in community question answering archives using a hierarchy of classifiers. Information Sciences, 261, (pp. 101-115).

Wang, Xin-Jing and Tu, Xudong and Feng, Dan and Zhang, Lei (2009). Ranking community answers by modeling question-answer relationships via analogical reasoning. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 179-186).