Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Answer Quality Evaluation in Online Health Care Community

Binjun Zhu[a], Xiaofeng Cai[b] and Ruichu Cai[c]

Faculty of Computer Science Guangdong University of TechnologyGuangzhou, China

[a] Corresponding author: zhubinjun1993@163.com
[b] ms.xiaofengcai@hotmail.com
[c] cairuichu@gmail.com

Abstract. Nowadays, through the online health care community platform, users can raise health related questions and doctors would provide corresponding answers. However, some answers could be low-quality and repeated. In this paper, we aim to evaluate and predict the quality of the answers in online health care communities. We set the evaluation rules and scoring model for the medical answer text. 12 features are proposed to represent the answer quality. 8 classic classification models are used to predict the answer score. The best model gets 0.90 f1-score. Furthermore, we utilize our model to select QA data of high quality, which help the QA matching task and promote the f1-score from 0.86 to 0.93.

Key words: evaluation rules; QA; promote the f1-score.

INTRODUCTION

The online medical QA websites provide a convenient and low-cost platform for the users to post their questions and get answers from doctors worldwide. To respond efficiently, doctor usually copy the most frequently used answers on the web to the patients. Those answers, replied by the professional doctors, are more reliable and rarely misleading. However, the quality of the answers is varying, and it would impact related NLP tasks. Doctors can copy their knowledge summary to reply, which have high value of utilization. But they might have to send the meaningless answer such as asking for more detail information and description, which have less medical knowledge.

It is really a challenge to evaluate the quality of medical answers. Firstly, the answer quality doesn't have an absolute and quantitative standard which is an uncertain work. For example, for such answer:" You can contact me with the phone number: xxx-xxxxxxx", if what the user required is the contact information of the doctor, it's a good answer. If not, the answer is just a spam without any useful medical knowledge. Both the amount of useful information and the matched degree of the QA pair should be considered into the answer quality evaluation. Moreover, the medical texts which contain lots of professional and complex concepts, is difficult to be understood by the computer. Therefore, appropriate features should be applied to represent the quality of answer.

In this paper, we propose a method to evaluate and predict the quality of medical answer on health QA websites. The answer quality is defined by the amount of medical information and matching degree of QA pair. We classify the answer texts into 3 levels and propose 3 evaluation rules. We also propose 12 features to represent the answer quality in the grouped QA data using the NER, Word2Vec models. 8 classification models are used to valid the effect of the features by our dataset. The best model gets 0.92 f1-score and 0.93 accuracy. Moreover, to verify the effect of our model, the model with best evaluation result would be used to select the high-quality QA data, which promote f1-score from 0.86 to 0.91 up about 15%.

MEDICAL ANSWER QUALITY EVALUATION

Evaluation Rule.

Although there are some research about medical text information quality evaluation [1][2], there is no definitive standard for medical answer quality evaluation yet. We propose 3 labeling standards after analysis and summary for the health community answers as bellow:

- Evaluate whether the answer is "meaningful" from user's perspective. It means we should consider whether patients feel helpful, not answer looks like correct towards questions from doctor's view. Patients always want to get more information directly about their disease (like cost, cure duration, the best treatment etc.). Especially we assume repeated answer is summary from doctor's experience.
- From top score (2 points) to tail score (0 point), look for absolutely negative cases in the candidate 5 questions to assign score. That may represent doctor's expertise in forum.
- Evaluate whether the answer can be applied on the candidate questions themselves and potential more similar questions. This is scalable to apply it to large scale medical QA system.

Answer Scoring

With the evaluation standard, the answer can be classified into 3 categories: meaningless, referenceable and valuable information with 0-2 grade score.

TABLE 1. Scoring for Medical Answer Quality

Score	Explanation	Example
0	meaningless information	"You should complement more information about your disease."
1	Referenceable information	"You should go to hospital and make the routine blood tests firstly."
2	valuable information	"You can use Metformin, which is the front-line, go-to treatment for diabetes."

Medical Answer Text Feature

Each item of QA data crawled from the health community has multiple data field, like: question text, answer text, disease name, and faculty name. We also categorize all the QA data which have same answer to build the Grouped QA Data. We propose 12 features for the representation of the copied answer group data. The features are composed by 3 categories:

TABLE 2. Medical Answer Quality Feature

Index	Feature Name	Description
1	disease_cnt	Disease entity count in answer text
2	symptom_cnt	Symptom entity count in answer text
3	medicine_cnt	Medicine entity count in answer text
4	surgery_cnt	Surgery entity count in answer text
5	examination_cnt	Examination entity count in answer text
6	body_cnt	Body entity count in answer text
7	entity_cnt	All kinds of entities count in answer text
8	w2v_ave	Average mean of all the word vectors from segmented answer text
9	answer_len	Length of Answer Text
10	uni_disease_count	Unique disease name in all the queries
11	uni_faculty_count	Unique faculty name in all the queries
12	question_count	Number of queries

- NER feature: The Stanford NER [3] model is trained with our manual labeled data, which can tag the 6 kinds of entity in the text: 1. Disease 2. Symtom 3. Surgery 4. Examination 5. Medicine 6. Body. Therefore, we can extract 1-7 features in Table 2.
- Embedding feature: All the questions and the answer texts in QA data are used for training the Word2Vec [4] model, which can extract the 8 features in Table 2.
- QA feather: Each copied answer group with multiple QA data can be extracted 9-12 features in Table 2.

Prediction Model

8 classic classification models are used for our evaluation task: 1. Logistic Regression (LR), 2. Naive Bayes (NB), 3.SVM with linear kernel (SVM_linear), 4.SVM with radial basis function kernel (SVM_rbf), 5. Decision Tree (DT), 6. Logistic Regression (LR), 7. Random Forest (RF), 8. Gradient Boosting Decision Tree (GBDT).

EXPERIMENT

Medical QA Corpus

We have 4,006,206 QA data crawled from the health website. After the QA data group with same answer, there are 73,778 copied answer group data, and each group have an average of 7.6 QA data. The repeated answers accounts for 13.96% of all the answers.

Answer Quality Prediction Task

Dataset

900 QA pairs are selected randomly from the groups which have more than 5 QA data. Our three annotators labeled the data in 0-2 scores. Finally, the answer quality dataset is built and have 375(0 score), 306(1 score) and 34(2 score) labels. The Stanford NER [3], Word2Vec [4] and Chinese tokenizing and POS tagging [5] models are used for extract features from the medical text.

Result

The 12 features are used to represent the answer quality. And the manual labeled data is used for 5-fold cross validation. Through the classification experiment, all the models get excellent result, which also verified the efficiency of the features. Eventually the random forest model gets the best performance: 0.921 f1-score and 0.931 accuracy.

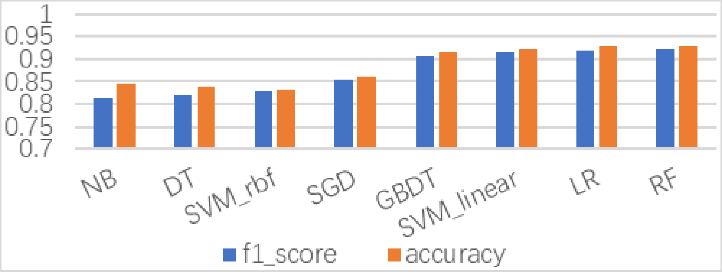

FIGURE 1. Answer Quality Evaluation Result.

In this paper, we propose a method to evaluate and predict the quality of medical answer in health QA website. The answer quality is defined by the amount of medical information and matching degree of QA pair. We classify the answers into 3 categories and propose 3 evaluation rules. We also propose 12 features to represent the answer quality in the grouped QA data using the NER, Word2Vec models. 8 classifiers are used to valid the effect of the features by our dataset. Moreover, the high-quality QA data, selected by the best evaluation model, have promoted the performance of the QA matching task.

QA Matching Task

Description

To validate the effect of high quality answer data (more than 1 score) predicted by our method, we try to use the high-quality data and random selected QA data to complete the QA matching task. We assume exist QA data are the positive samples, and randomly allocate an answer for the question in positive samples to build the negative samples.

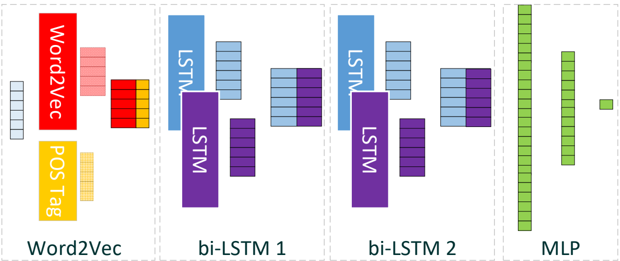

FIGURE 2. Deep Neural Network for QA matching Task.

We use a deep neural network for the QA matching task [6]. We replace the DNN model in the paper by the structure in Figure 2. An extract POS Tagging model is added in the first Word Embedding layers. So, each vector of word has both the semantic and POS information. Then we apply 2 Bidirectional LSTM [7][8] layers to extract semantic feature. Finally, 2 full-connected layers are used to reduce the feature dimension and predict the final result.

Result

Both the random QA dataset and the high-quality QA dataset have 10,000 positive and 10,000 negative samples. We split the dataset in 60% training data, 20% valid data, 20% test data. The result of high quality dataset has obvious promotion, which of the best f1-score reached 0.91, up by more than 15% on the best f1-score 0.86 in random QA dataset.

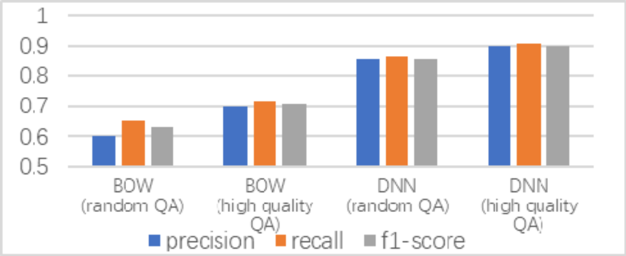

FIGURE 3. Result of QA matching task and improvement.

CONCLUSION

This paper presents 12 medical answer qualities for answer quality evaluation, which are extracted by medical NER model, word2vec model, and manual design. The cross-validation experiment results verify the effectiveness of the proposed method of answer quality evaluation. With our method, the high-quality QA data also helps the QA matching task and improve the baseline result.

REFERENCES

1. Oh, S., Yi, Y. J., & Worrall, A. (2012). Quality of health answers in social Q&A. Proceedings of the Association for Information Science and Technology, 49(1), 1-6.
2. Oh, S., & Worrall, A. (2013). Health answer quality evaluation by librarians, nurses, and users in social Q&A. Library & information science research, 35(4), 288-298.
3. Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 363-370). Association for Computational Linguistics.
4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
5. Zhang, H. P., Yu, H. K., Xiong, D. Y., & Liu, Q. (2003, July). HHMM-based Chinese lexical analyzer ICTCLAS. In Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17 (pp. 184-187). Association for Computational Linguistics.
6. Feng, M., Xiang, B., Glass, M. R., Wang, L., & Zhou, B. (2015, December). Applying deep learning to answer selection: A study and an open task. In Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on (pp. 813-820). IEEE.
7. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
8. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6), 602-610.