Discussion on The Data Extraction Strategy Of FAQ System Based On Chat Records

Jia Li¹*, Xin Chen²

¹Department of Computer Science, Xiamen Institute of Technology, XiaMen, FuJian, 361021, China
²Xiamen Kuaishangtong Technology Co., Ltd, XiaMen, FuJian,361000, China
*Corresponding author’s e-mail:lijia30090@yeah.net

Abstract. Q&A data based on chat records have one characteristic: customers ask ,service answers. There is a large amount of knowledge between the questions and answers of customer service. By optimizing the Q&A extraction algorithm to extract knowledge, a very excellent Q&A library can be constructed, thus the accuracy of FAQ system is greatly improved. By analyzing the existing data, this paper considers the extraction strategy of question answering from machine learning and non-machine learning respectively. Then We compares their performance from three aspects of precision, recall and F1-score according to their different characteristics. In order to ensure the best classification performance, grid search and K-fold crossover are also used to test the optimized classifier performance. After selecting the optimal data extraction strategy,We developed a FAQ system use this strategy, the system results show that the performance is reliable.

1. Introduction
In FAQ system, data is the basis of the system, and its quality often determines the answer effect of the FAQ system[1]. Traditional search engines can only return documents or web pages that match keywords, while question-answering system based on natural language processing technology can improve the mode of returning search results of traditional search engines. The FAQ system itself inputs questions in natural language. By analyzing the intent of the user’s question, the corresponding answer is matched in the Q&A library according to the intent and returned to the user in natural language, instead of returning a large number of documents or web pages, this has greatly improved both in user experience and time cost. Q&A extraction is the most important step in developing FAQ system[2].

The rest of this paper describes the relevant models and evaluation indicators involved in data extraction at first. then We focus on data extraction strategies. The core work of data extraction is to distinguish question from answer[3]. In order to find a suitable classification algorithm, a total of four models of support vector machines, naive Bayes, decision trees and regular expressions are compared to learn from sample data [4]. Then grid search and K-fold cross are checked to what extent the optimized model can reach, so that we can determine whether it is necessary to use integrated learning to combine the model. Finally, the paper shows the presentation effect of the FAQ system based on chat records briefly.
2. Relevant models and evaluation indicators

2.1 Word2Vec
Word2vec is essentially a simplistic neural network. Word2vec specifically refers to CBOW and skip-gram. The Skip-Gram model uses the prediction results of surrounding words and uses gradient descent to adjust the word vector of the central word continuously. After all the text is traversed, the word vector of all words in the text is obtained. Compared with CBOW [5], Skip-Gram uses surrounding words to constrain the current words. When the amount of data is sufficient (without considering the amount of training calculation), in theory, the effect of Skip-Gram will be better than CBOW.

2.2 Grid search and K-fold cross
Grid search and K-fold cross can maximize the use of data and optimize the classifier. By looping through the set parameters, trying every possibility, and finally getting the best result, it is gradually accepted by people in terms of parameter tuning. The advantage of K-fold cross-checking is that it can ensure that every piece of data has a chance to be trained and predicted, and it also makes the performance of the optimized classifier more reliable as much as possible [6] (in the following experiments, the K value is 10). However, both of them have an obvious shortcoming that is slow. Generally, Using these two methods together, the optimal classifier can be obtained when training the classifier without too much human intervention in the process.

2.3 Integrated learning
Integrated learning is also called multi-classifier system [7], which achieves the purpose of learning tasks by constructing and combining multiple learning. There are three types of integrated learning. The first method is the learning method, the typical representative is the Stacking algorithm. Stacking first trains a primary learner from the initial training set. The output data of this primary learner is used as the input data of the secondary learner, and the label of the initial sample is still used as the sample label. Stacking is often used with K-fold cross. The second method is Boosting. The Boosting family is an algorithm that can promote a weak learner to a strong learner. Its algorithm principle is: first train a learner 0 with training set 0, then adjust the distribution of training samples according to the performance of learner 0, Thus the training samples that learn wrongly in learner 0 receive more attention in the subsequent learning, and then train the next learner t according to the adjusted sample distribution; and so on, until the number of learner learning reaches the initially set value T, Finally, combine the T base learners by weight. The typical learner is XGBoost. The third method is the average method, which is mostly used for regression. According to the idea of integrated learning, for all classifiers, each classifier has its own specific sentence pattern that is good at classifying. In the case of poor training index, using the result of voting by an odd number of classifiers, according to the principle of minority obeying the majority, Integrate a classifier as the final output.

2.4 precision and recall
Precision, also known as precision, It represents the proportion of positive cases in which positive cases are divided into positive cases. The calculation formula is TP/(TP+FP). For example, in this chapter, Precision of question= the number of questions that the classifier predicted to be positive and actually were positive/the total number of questions. and the Precision of declarative sentences is calculated in the same way.

Recall is also called recall rate, which represents the probability of being predicted as a positive sample in a sample that is actually whole. The calculation formula is TP/(TP+FN). For example, in this chapter, Recall of question = the number of questions that the classifier predicted to be positive and actually were positive / (the number of questions that the classifier predicted to be positive and actually were positive + the number of questions that the classifier predicted to be positive and actually were negative). the recall of declarative sentences is calculated in the same way.
2.5 f1-score
The evaluation index also uses f1-score. f1-score is a weighted average that takes into account both precision and recall. Its maximum value is 1 and its minimum value is 0. The calculation formula is \( f1\)-score\(=\frac{2\times (\text{precision}\times \text{recall})}{\text{precision} + \text{recall}} \). The purpose of using f1-score is to avoid the extreme situation where one of the indicators of precision or recall is 0 and the other is 1.

3. Q&A extraction strategy
Q&A extraction is divided into question extraction and answer extraction. The premise of extracting Q&A is to identify all the questions and find the answers one by one. If it is the other way, it is difficult to find the corresponding question. The sample data is shown in Table 1.

| Id | Role  | Sentence                           |
|----|-------|------------------------------------|
| 0  | Client| How to get a credit card?          |
| 1  | Server| Hello, you can hold the ID card at the bank counter. |
| 2  | Server| Can also be handled online through mobile banking. |
| 3  | Server| Do you need to do that?            |
| 4  | Client| Yes                                |

It should be noted that valid Q&A pairs need to meet client questions and server answers. In Table 1, the valid Q&A pair is \([0, 1]\). If the server asks a question, the client answer is not counted as a valid Q&A pair, because the reply role in the Q&A system is the server.

3.1 Selection of Q&A classification algorithm
In order to find a suitable classifier, four classification algorithms: support vector machine, naive Bayes, decision tree, and regular expression are compared to learn sample data. The sample data set is the data about the financial field crawled from Zhihu by the crawler. There are 6000 pieces of data in the training set and the test set (input in the form of QA, the ratio of questions and answers is 1:1). The whole data set is divided according to the ratio of training: test is equal to 8:2.

| SVM | NBM | DTM | RE  |
|-----|-----|-----|-----|
| Precision |
| 0    | 0.96| 0.85| 0.85| 0.94|
| 1    | 0.95| 0.88| 0.85| 0.93|
| Avg/total | 0.96 | 0.86 | 0.85 | 0.94 |
| Recall |
| 0    | 0.96| 0.9  | 0.87 | 0.92 |
| 1    | 0.96| 0.82 | 0.83 | 0.94 |
| Avg/total | 0.96 | 0.86 | 0.85 | 0.94 |
| F1-score |
| 0    | 0.96| 0.87 | 0.86 | 0.94 |
| 1    | 0.96| 0.85 | 0.84 | 0.94 |
| Avg/total | 0.96 | 0.86 | 0.85 | 0.94 |

Table 2 shows the comparison of support vector machines, naive Bayes, decision trees, and regular expressions in question recognition. The default parameters of the support vector machine is\( C=1.0, \) cache\_size=200, class\_weight=None, coef0=0.0, decision\_function\_shape='o vr', degree=3, gamma='auto', kernel='rbf', max\_iter=-1, probability=False, random\_state=None, shrinking=True, tol=0.001, verbose=False. The default parameter for naive bayes is priors=None. The default parameter for decision tree is\( \text{class}weight=None, \text{criterion}='\text{gini}', \text{max}dept=h=\text{None}, \text{max}features=None, \text{max}leaf\_nodes=None, \text{min}impurity\_decrease=0.0, \text{min}impurity\_split=\text{None}, \text{min}samples\_leaf=1, \text{min}samples\_split=2, \text{min}weight\_fraction\_leaf=0.0, \text{pre}sort=False, \text{random}state=None, \text{splitter}='best'. \)
As can be seen from Table 2, support vector machine (SVM) is better than naive Bayes and decision tree in terms of precision, recall and F1 when processing text tasks, while regular expression is slightly weaker than SVM and better than decision tree and naive Bayes. The three classifiers of support vector machine, decision tree, and naive Bayes are all trained with default parameters. Compared with the default parameters, manually setting the parameters may be better in results. Since the regular expression "parameter adjustment" method is to do regression testing after manual accumulation and adding rules, in terms of efficiency, the effect of regular expressions is the best foreseen, but there are bottlenecks in optimization. In the later, if you want to increase by 1%, you need a large amount of data to study. Especially, for some sentences without question keywords, the recognition effect is poor. Machine learning algorithms can just overcome it.

In view of the fact that the three classifiers of support vector machine, naive Bayes, and decision tree are not adjusted, there may be insufficient data volume, which will affect the results of the classifier itself. Therefore, the semi-supervised algorithm tri-training method \cite{8} is preferred here to expand the training set and test set. It is worth mentioning that using this method can greatly save labeling time.

3.2 Optimize classifier

Before using the Tri-Training method, 3000 questions and 3000 answers were manually labeled on the data set. Plus 6000 question and answer sentences in the training set and test set mentioned above, a total of 12000 question and answer sentences were labeled. The ratio of sentences to answers is equal to 1:1. After using the tri-training semi-supervised algorithm, more than 9000 questions and 9000 answers are extracted. After re-weighting and manual cleaning, there are about 30000 questions and answers left, among which the ratio of questions and answers is 1:1. Next, the classifier is optimized by tuning parameters on this data set, and the performance of the optimized classifier is tested by grid search and K-fold crossover, so as to clarify whether Integrated learning is needed to further combine the classifier to ensure the best classification effect.

After the classifier was optimized by grid search and 10-fold cross test, the relevant indexes were shown in Figure 1. SVM stands for Support Vector Machine, NBM stands for Naive Bayes, and DTM stands for Decision Tree, Re represents a regular expression, and the optimized SVM classifier parameters are as follows: SVC(C=4, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma=0.5, kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, Verbose = False). The parameters of the optimized Naive Bayes classifier are as follows: Gaussiannb(Priors=[0.02, 0.98]). DecisionTreeClassifier (class_weight=None, criterion='gini', max_depth=9, Max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best').
As can be seen from Figure 1, the average accuracy, recall rate and F1-score of the SVM classifier are all increased by 1%, the average accuracy, recall rate and F1-score of the Naive Bayes classifier are all increased by 1%, and the average accuracy, recall rate and F1-score of the decision tree classifier are all increased by 5%. However, SVM is superior to Naive Bayes, Decision Tree and Regular Expression in terms of accuracy, recall rate and F1-score.

For the optimized classifier, the highest SVM has reached an average accuracy of 0.97. For integrated learning, it integrates multiple classifiers, so the running time is at least equal to the sum of the time of these primary classifiers. After multiple statistics, Running time of the four algorithms is shown in Table 3.

| Time  | Support Vector Machine | Naive Bayes | Decision Tree | Regular Expression |
|-------|------------------------|-------------|---------------|---------------------|
|       | 22.67                  | 0.015       | 0.402         | 0.001               |

It can be clearly seen from Table 3 that the running time of support vector machine is much longer than other three algorithms, and the running speed of the other three algorithms is negligible compared with SVM. Comprehensive performance and time, the three algorithms with the best indicators (support vector machine, naive Bayes and regular expression) are selected. The running time after using integrated learning is 24.002ms. The test indicators are shown in Table 4.

In Table 4, after using Integrated learning (voting method), the integrated classifier (a classifier that integrates svm, naive Bayes and regular expression classification results for voting) is only 0.01 better than SVM, and the time consumption is 1.3 milliseconds longer. This is not a big deal for predicting one item, but when the data to be predicted reaches more than one million level, the time will be more than tens of minutes. Therefore, considering the time dimension and performance, Integrated learning is abandoned and support vector machine is preferred for question recognition.

| True          | Precision | Recall | F1-score | Support | True  |
|---------------|-----------|--------|----------|---------|-------|
| 0             | 0.98      | 0.97   | 0.98     | 2963    | 2904  |
| 1             | 0.97      | 0.98   | 0.98     | 3037    | 2946  |
| **Avg/total** | **0.98**  | **0.98** | **0.98** | **6000** | **-** |
After identifying and classifying all the questions and answers in the data set, the next work is to match questions to answers. In our systems we only consider the form of one question one answer and one question two (multiple) answers \[^9\].

4. Results Presentation

After the FAQ data set is constructed, in order to present the effect of the FAQ system, we need to match the questions input by users and present the results to them. At this stage, similarity matching is mainly considered\[^10\]. We set threshold value is 0.8, and select the top three questions with the highest similarity as preselection questions. Then, compare the length of preselection questions and user input questions, the answer to the question whose length is closest to the preselection question is returned to the user. If there are no questions greater than the threshold, the user is prompted that there is no similar answer to the current input question. The results of system shows that the effect is good.

5. Conclusion

In this paper, We discuss the data extraction strategy of FAQ system based on chat records. Since the data set is financial data crawled from Zhihu, choosing a suitable classification algorithm will have a great impact on the results. We consider the extraction strategy of QA from machine learning and non-machine learning respectively. Considering all aspects, the results presented by the system are good. However, there are still many details that can be further optimized. For example, some statements may also be problems, which can be included in the training data to improve the performance of the system. You can also fully consider the context of the entire dialogue\[^11\], "remember" more details, thus improving the accuracy of the system.

Acknowledgments

This work has been supported by the Fujian Province Young and Middle-aged Teacher Education Research Project (Natural Science), the project No. is JAT200830.

References

[1] Bo Zheng, Haoyang Wen etc. (2020) Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension. In: ACL.Seattle(OnLine).
[2] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. (2019) Synthetic QA corpora generation with roundtrip consistency. In ACL.Florence.
[3] Wu Y, Wu W, Xing C, Xu C, Li Z, Zhou M.(2019) A sequential matching framework for multi-turn response selection in retrieval-based chatbots. Comput Linguist.45(1):163–97.
[4] Kenter, Tom.(2016) Siamese CBOV: Optimizing Word Embeddings for Sentence Representations. In: ACL. Berlin.
[5] Li, H. (2012) Statistical Learning Methods. Tsinghua University Press, Beijing.
[6] Jung, Y. (2017) Multiple predicting K-fold cross-validation for model selection. Journal of Nonparametric Statistics.1-19.
[7] Qiu A.K., zhu J.G. (2017) Multi-set canonical correlation analysis method based on Inte- ted learning. Computer Engineering and Applications.53(6):162-169.
[8] Qian T.Y., Liu B., Chen L., Peng Z.Y.(2014) Tri-Training for authorship attribution with limited training data.In:ACL.Baltimore.pp.345-351.
[9] Karan M., Snajder J.(2018) Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. Expert Systems with Applications, 91:418-433.
[10] Sakata W., Shibata T., Tanaka R.,Kurohashi S.(2019) FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance. In:ACM. Paris. pp. 1113 -1116.
[11] Tan M., Santos C. D., Xiang B.(2016) LSTM-based Deep Learning Models for Non-facto id Answer Selection. In:ICLR. San juan.