Incorporating Domain Knowledge into Text Classification Diagnosis in Customer Service Dialogue Field

Jiangjiang Zhao¹, Jie Zhu², *, Xiaokun Zhang¹, Xian-Ling Mao² and Heyan Huang²

¹ China Mobile Online Services Co., Ltd, Beijing, China
² Beijing Institute of Technology, Beijing, China

*Email: zhu_j@bit.edu.cn

Abstract. The customer service dialogue process is an important way for consumers to communicate with manufacturers. In order to enhance the consumer experience as well as to assist the staff, we build a knowledge base that can categorize consumer questions and provide suitable answers. However, due to labeling deviations, there are some errors in the knowledge base. So we propose a domain knowledge-based text classification diagnosis method, which innovatively transforms the question and answer task into the text classification task. We use an ERNIE-based structure to match consumer questions with multivariate groups of answers from the knowledge base, judged by similarity. Also for incorrectly matched pairs, our method provides a list of suitable candidates for selection. Compared with other baselines, our model achieves competitive results. At the same time, good results are obtained on cross-province data, proving that our method has good scalability.

1. Introduction

In the communication field, there is a huge amount of customer service conversation logs, which contain much important information, such as customer requests, product reviews, marketing opportunities and so on. Nowadays, we offer our services in dozens of provinces and collect a lot of available user log information. Due to the high training cost and error-prone of customer service staff, and the limited working capabilities of online robots, we build a knowledge base system to assist them in their work. Taking into account regional differences, our branches in different provinces build and maintain a set of knowledge base system with more than 3000 scenarios (also called standard answers), and 80,000 to 100,000 user questions.

We build a system based on domain-knowledge to assist the staff, and efficiently solve the questions raised by the users in the process of customer service, that is automatically obtain question-answer pairs from a given corpus. The main challenge of automatic question and answer mining is the lexical gap between query questions and historical questions [1]. [2] proposes a classification method based on sequential patterns to detect questions in forum topics, and proposes a graph-based propagation method to detect answers to questions in the same topic. [3] captures the complex semantic relationship between the question and the answer through a hybrid model. [4] proposes a new method to use question category information to improve the performance of question retrieval. In recent years, the generation method has attracted more and more attention. Previous work has proposed a variety of question generation methods to improve the existing question answering system [5]. However, previous work has many limitations in this task. First, these methods cannot make full use of the unique domain knowledge in
our knowledge base. Second, the generation method cannot meet the implementation standards in terms of generation quality and fluency. At the same time, online customer service has high requirements for responses.

In order to avoid the above limitations, we innovatively regard the automatic question and answer mining task as a text classification task. Given a similar question and a standard answer pair, we judge whether the relationship between the pair matches. To solve the matching problem of standard answers and similar problems, we use a variety of BERT-based pre-training models to perform linear transformation on the [CLS] representation, and apply the softmax function on it for prediction. However, in the application, we find that there are some errors or deviations in the manually constructed matching relationship. For example, the standard answer\textsubscript{1} corresponds to the similar problem\textsubscript{1}, but it is not as reasonable as the matching standard answer\textsubscript{2}. Therefore, we also use text classification methods to diagnose the knowledge base system. At the same time, we can automatically collect new problem pairs by finding the relationship between existing unsolved similar problems and standard answers, further expanding the content of the knowledge base and improving its quality.

Our main contributions are as follows:

- By summarizing the customer service dialogue logs, we propose a brand-new knowledge base that contains more than 4000 standard answers and 80,000 similar questions. At the same time, we manually annotate the customer service dialogue logs, and obtain more than 40,000 matching pairs of similar questions and standard answers.

- With the aid of domain knowledge, we innovatively regard the automatic question and answer mining task as a text classification task, and propose an applicable method based on ERNIE.

- Considering the cross-province application, we transfer the model to other provinces for data testing and obtain good results, which proves the effectiveness and robustness of our method.

2. Related Work

2.1 Text Classification

Text classification is a classical task in natural language processing (NLP), which aims to assign labels to a text such as sentences, queries, paragraphs, and documents. Text classification in machine learning based approaches learns to make classifications based on training data. Early feed-forward neural network [6] focuses on assign the bag-of-word representation of a text into a certain category. RNN-based models [7] regard the text as a sequence and CNN-based models [8] aim to recognize the pattern (such as key phrases) in the text. Transformer-based models [9] make it possible to pre-train a large language model for classification.

2.2 Pre-train Model

The rise of large-scale pre-trained models improves the performance of many NLP tasks. Transformer-based pre-trained models such as OpenGPT [10,11] and BERT [12] have much deeper network architectures and pre-train on more training data than previous CNN or LSTM-based contextualized embedding models [13,14]. It makes transformer-based models competitive and popular in recent years.

3. Method

To tackle the matching problem, we leverage the ERNIE [15] pretrained model. We implement a linear transformation to the representation of [CLS] and apply a softmax function on it to make the prediction. Specifically, the input of our problem is \( X = (x_1, x_2, ..., x_n, x'_1, ..., x'_m) \), where \( x_i (1 \leq i \leq n) \) is the \( i^{th} \) token of the similar question, while \( x'_j (1 \leq j \leq m) \) is the \( j^{th} \) token of the standard answer. For each token, we sum its word embedding as well as its positional embedding as the token representation, and feed it to the model with the other token representations.

To predict whether the question pair matches, we use the follow equation to calculate the probability:

\[
p = \text{softmax} \left( C \cdot W^T + b \right),
\]  
(1)
where \( C \in \mathbb{R}^H \) is the representation of [CLS] (\( H \) is the dimension of the hidden vector), \( W \in \mathbb{R}^{K \times H} \) is the linear transformation of \( C \) (\( K = 2 \) is the number of classes, i.e. matched or unmatched) and \( p \in \mathbb{R}^K \) is the probabilities of each class. The architecture of the model is shown in Figure 1.

![Figure 1. The architecture of our model.](image)

4. Knowledge Base

4.1 Data Collection

Our task is to conduct data mining from the current massive customer service conversation logs and build a system based on domain knowledge. The system can be used to assist customer service personnel and online robots to complete their work and reduce costs. As the relevant data of Henan Province is relatively complete, we conduct data analysis based on the Henan Province Dialogue Log. We construct a standard answer base and a user question base respectively. The standard answer base is used to classify the scenes that appear in the customer log, and the user question knowledge base stores the matching relationship between typical user questions and standard answers.

After the construction of the knowledge base system, we find some errors and deviations in the stored knowledge information during usage. This brings some interference to the work. So we extract the knowledge base data, splice user questions and standard answers, and label them for error correction.

4.2 Data Properties

The standard answer library mainly describes the various situations of customer service scenarios and some of the corresponding attributes. The field contents are as follows:

- **Standard_answer.** This field is used to describe customer service scenarios.
- **Keywords.** This field contains several words, which are strongly related to standard answers.
- **Category_name.** This field describes detailed categories, including three levels.
- **Answer.** This field represents the detailed content of standard answers.

The user question knowledge base focuses on describing the templates of questions customers may ask and the corresponding standard categories of questions. The field contents are as follows:

- **User_question.** This field describes some possible questions from users.
- **Standard_answer.** This field is most suitable standard answer for user question, and also labelled by experts.
- **Category_name.** This field describes detailed categories corresponding to the standard answer.
- **Keywords.** This field describes special keywords corresponding to the standard answer.

In practice, we consider that the core of the problem lies in diagnosing the wrong matching relationship, so we process the data into the form of \(<\text{User_question, Standard_answer, Label, Attribute}>\).
4.3 Information Statistic
We annotate 40,000 multivariate data sets with a 10:1 ratio of positive to negative, in which all correct use cases are annotated manually, a few incorrect use cases are annotated manually, and the rest are generated by matching existing relationships and verified manually. When selecting data, we need to ensure that a minimum of 5 of each type of data exists, and then select as many different types of data as possible.

Our statistics on the relevant attributes of the data yield the following results.
- Basic data length information, as shown in Figure 2. Statistically, the interval for the length of the dialogue text is [2, 118], with a mean of 19.0058, a median of 18, and a plurality of 16.
- A message that matches the wrong data. Statistically, 12.5% of the labelled data are incorrect.
- There are five main types of errors, of which 37% are “Misspelt or Mispronounced characters”, 23% are “Ambiguity of meaning or Relationships”, 20% are “Same type but different meaning”, 11% are “Business Consulting” and 9% are “Other types”.

Figure 2. Frequency distribution of number of sentences.

5. Experiments

5.1 Main Results
We use four baseline methods in the following experiment:

TextCNN: TextCNN [16] is a CNN-based model used in sentence-level text classification tasks. In this baseline, we feed the concatenation of standard answer and similar question representations into the TextCNN model and the output is the probability of the match of the question pair.

BERT+MLP: In this baseline, the sentence embedding representation of the BERT [8] is the input of an MLP module. The softmax function is applied to the output of the MLP module, which predict whether the question pair is matched.

BERT-base: In this baseline, the output of the [CLS] representation of the BERT is the input of an MLP module. Same as before, the softmax function is applied to the output of the MLP module, which predict whether the question pair is matched.

BERT+DPCNN: DPCNN [17] is a CNN-based model using word-level deep pyramid, which can effectively extract the remote relation features in text. Like the BERT-base mentioned above, after obtaining the representation of the [CLS] token from the BERT module, it is fed into the DPCNN module with the softmax function.

We use three metrics to evaluate our model: Accuracy, Recall and F1 score. For the testing phase, the result of the example can be one of true positive (TP), true negative (TN), false negative (FN) and false positive (FP). The definitions of Accuracy, Recall and F1 score are as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TP + TN} \quad (2)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (3)
\]
The results into the model, sort them according to the probability value, and take out the top five largest numbers. We manually select from the list, choose suitable options, and modify the content in the existing knowledge base to correct errors. Some examples are listed in Table 3.

After confirming the effect of our model, the next step is to correct the mis-matched sentence pairs. For the wrong matching sentences in the test set, we take out similar questions and match them with all other standard answers one by one. Then, we send the results into the model, sort them according to the probability value, and take out the top five largest numbers. We manually select from the list, choose the most suitable option and modify the content in the existing knowledge base to correct errors. Some examples are listed in Table 3.

For example, in the first line, the user question "I would like to inquire about group ringtone." means that the customer wants to know some information about group ringtone, but not shows that he would like to subscribe this service directly. The previous error standard answer "Group ringtone subscription" ignores the content of the user's meaning of "inquire", which is clearly inappropriate. In comparison, "Introduction to group ringtone" in candidate Standard Answers list is more suitable.

The candidate lists in the latter examples also contain the most appropriate relationships, indicating that our model also has excellent ability in finding the correct relationship matches.

\[
F1 \text{ score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

Table 1 shows Accuracy, Recall and F1 scores of our model and the baseline models on the test set.

Table 1. The performances of our model and the baseline models.

| Models    | Accuracy | Recall | F1 score |
|-----------|----------|--------|----------|
| TextCNN   | 0.6598   | 0.6643 | 0.7679   |
| BERT+MLP  | 0.7294   | 0.8608 | 0.8136   |
| BERT-base | 0.8136   | 0.8402 | 0.8929   |
| BERT+DPCNN| 0.8241   | 0.8851 | 0.8900   |
| Ours      | 0.8869   | 0.9582 | 0.9391   |

From the result in Table 1, we can see that the performance of our model is significantly better than the other four baseline methods in Accuracy, Recall and F1 score. Therefore, our model is competitive.

5.2 Generalization Study between Provinces

In mobile communications, business halls of different provinces may offer completely different packages or plans for their users. However, the content of the database in each province is not allowed to be shared for privacy reasons, making the generalization of our model important. In this part of our experience, we would like to examine how our model is generalized to different provinces.

We test our model on test data sets from four provinces. Each test data set has hundreds of question pairs with labels, including 1001 for Henan, 596 for Zhejiang, 736 for Shandong, and 621 for Yunnan. We train our model as well as other baseline models on the Henan dataset, and compare their performance by Accuracy, Recall, and F1 score on other test sets.

As Table 2 shows, our method far surpasses all benchmark models for Accuracy and F1 score. We change the test data from Henan Province to other provinces, and find that the various evaluation indicators remain stable and have improved. This shows that our method presents the best stability and generalization, and has a wide and efficient application value.

Table 2. Generalization study results between provinces.

| Model   | Henan Acc | Zhejiang Acc | Yunnan Acc | Shandong Acc | Henan F1 | Zhejiang F1 | Yunnan F1 | Shandong F1 |
|---------|-----------|--------------|------------|--------------|----------|-------------|-----------|-------------|
| TextCNN | 66.43     | 73.29        | 79.26      | 64.60        | 74.24    | 70.21       | 79.47     | 62.91       |
| BERT+MLP| 71.43     | 75.52        | 80.91      | 69.67        | 78.01    | 78.17       | 90.87     | 84.94       |
| BERT-base| 84.11    | 92.24        | 91.04      | 82.72        | 88.57    | 89.21       | 93.18     | 76.77       |
| BERT+DPCNN| 82.82   | 92.46        | 90.40      | 83.20        | 86.74    | 82.24       | 86.75     | 77.93       |
| Ours    | 86.31     | 97.60        | 92.58      | 96.20        | 96.20    | 98.06       | 96.74     | 93.94       | 85.78 | 92.32 | 92.32 |
| User Questions                                      | Error Standard Answers                  | Candidate Standard Answers List                                                                 |
|----------------------------------------------------|----------------------------------------|------------------------------------------------------------------------------------------------|
| I would like to inquire about group ringtone.       | Group ringtone subscription            | Cancel group ringtone<br>Introduction to group ringtone<br>Open group ringtone<br>Open ringtone<br>Open multimedia messaging service |
| How to add a sub-parent number to Safe-Home-School service? | Safe-Home-School service subscription | Safe-Home-School service cancellation<br>Safe-Home-School service subscription<br>Introduction to Safe-Home-School service<br>Function introduction to Safe-Home-School service<br>Introduction to Safe-Home service |
| Can I change my home broadband?                     | Introduction to broadband speed up     | Broadband cancellation<br>New broadband installation<br>Broadband activities<br>Business broadband service change<br>Broadband application |

Notes: “User Question” represents the questions asked by consumers, “Error Standard answers” represents the wrong relationships previously labelled by the staff, and “Candidate Standard Answer List” represents the list of relationships with the highest matching probability selected by our model. In the “Candidate Standard Answers List” column, the bolded ones indicate the best matches selected by the staff. All fields are from the knowledge base mentioned in chapter 4.

6. Conclusion
In response to the questions raised by consumers, we construct a question-and-answer pair knowledge base to assist customer service agents. Since some wrong annotation relations are found in the application, we propose a text classification diagnosis method based on domain knowledge. This method innovatively transforms the question-and-answer problem into a text classification task and obtains good results. At the same time, we use the model to re-match the sentences with incorrect matching relations, so that they can become correct matching relations.

Acknowledgements
The work is supported by The National Key Research and Development Program of China: 2018YFC0830705.

References
[1] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. Question-answer topic model for question retrieval in community question answering. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 2471–2474, 2012.
[2] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. Finding question-answer pairs from online forums. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 467–474, 2008.
[3] Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. Improved representation learning for question-answer matching. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 464–473, 2016.
[4] Xin Cao, Gao Cong, Bin Cui, and Christian S Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In Proceedings of the 19th international conference on World wide web, pages 201–210, 2010.
[5] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 866–874, 2017.
[6] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivalssyntactic methods for text classification. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), pages 1681–1691, 2015.

[7] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075, 2015.

[8] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.

[9] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In China National Conference on Chinese Computational Linguistics, pages 194–206. Springer, 2019.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[13] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. Journal of machine learning research, 12(ARTICLE):2493–2537, 2011.

[14] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

[15] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223, 2019.

[16] Yoon Kim. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[17] Rie Johnson and Tong Zhang. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 562–570, 2017.