A Model Ensemble Approach for Conversational Quadruple Extraction

Zijian Tu, Bo Zhang, Chuchu Jiang, Jian Wang(✉), and Hongfei Lin

School of Computer Science and Technology,
Dalian University of Technology, Dalian 116024, Liaoning, China
{tuzj,zhangbo1998,chuchu}@mail.dlut.edu.cn, {wangjian,hflin}@dlut.edu.cn

Abstract. Fine-grained sentiment analysis of dialogue text is crucial for the model to understand the conversational participants' viewpoints and provide accurate responses in generating replies. Unfortunately, in the field of conversational opinion mining, coarse-grained dialogue emotion analysis remains the mainstream approach, despite being unable to meet the actual needs in some specific scenarios such as customer service question and answer system. This work focuses on conversational aspect-based sentiment quadruple analysis, which aims to detect the sentiment quadruple of target-aspect-opinion-sentiment in a dialogue. In this study, we mainly extract triplets and judge the unique sentiment, which is determined by the target and opinion terms together. For this purpose, we fine-tune the pre-trained language models using the DiaASQ dataset. We optimize the rotation positional information embedding by combining the actual length of the dialogue text and use adversarial training to enhance the model's performance and robustness. Finally, We use beam search ensemble algorithm to improve the entire triplet extraction system's performance. Our system achieved an average F1 score 40.50 that ranked second in the Chinese dataset and fifth in the general dataset for the Conversational Aspect-based Sentiment Quadruple Analysis shared task at NLPCC-2023.

Keywords: conversational quadruple extraction · fine-grained sentiment analysis · beam search ensemble algorithm

1 Introduction

Aspect-based sentiment analysis has become a popular technique in natural language processing to identify people's opinions and attitudes towards products or services by analyzing the sentiment of the text [1]. However, traditional sentiment analysis techniques often fail to capture the conversational flow of a dialogue or conversation. Conversations are complex and dynamic, where different speakers may have different viewpoints and emotions towards distinct aspects of the target, making it crucial to perform fine-grained sentiment analysis of dialogue text. In certain specific scenarios, such as a customer service question

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 174–184, 2023.
https://doi.org/10.1007/978-3-031-44699-3_16

and answer system, it is insufficient to solely identify the emotions expressed by consumers during a conversation. It becomes more important to identify the specific viewpoints of consumers regarding different aspects of the product in order to effectively address post-sales issues and provide better assistance. However, coarse-grained dialogue emotion analysis remains the mainstream approach in the field of conversational opinion mining. It appears that despite incorporating dialogue into the fine-grained sentiment analysis (DiaASQ) [2], the model's actual performance is still subpar.

Our work focuses on conversational aspect-based sentiment quadruple analysis (CASQA), which aims to detect the sentiment quadruple of target-aspect-opinion-sentiment in a dialogue. As shown in Fig. 1, our task involves extracting triples such as 'Apple','power consumption' and 'can't hold' from multiple rounds of dialogue among four speakers who are discussing the various aspects of the iPhone's performance. The extracted triples like 'Apple', 'power consumption' and 'can't hold' will then be evaluated for negative emotional polarity. The Corresponding aspect-based quadruples extracted in this dialogue fragment are shown in the Table 1. CASQA enables us to identify sentiment with respect to the specific aspects and opinions expressed in the conversation and provides a more accurate understanding of the conversation's sentiment.

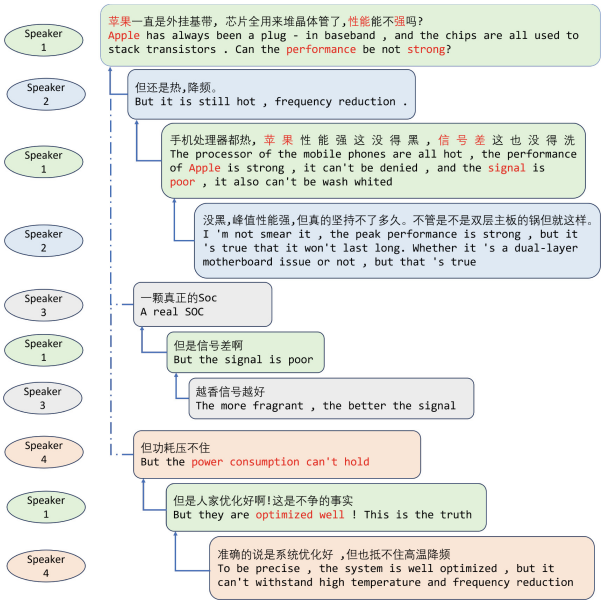

Fig. 1. Illustration of CASQA task

On the basis of the model and data set provided by DiaASQ, We achieved CASQA task by extracting triplets that judge the unique sentiment, which is determined by the target and opinion terms together.

Table 1. Corresponding aspect-based quadruples

Target	Aspect	Opinion	Sentiment
Apple	performance	strong	pos
Apple	signal	poor	neg
Apple	optimized	well	pos
Apple	power consumption	can't hold	neg

Our method is divided into two stages: (1) We first utilized a Neural Network Intelligence tool to search for hyperparameters that would lead to optimal performance of the model. Then we optimize the rotation positional information embedding (Roformer) [3] by combining the actual length of the dialogue text. Based on the discovered hyper-parameters, we fine-tuned the model using the Chinese-English dataset provided by DiaASQ. After gradient back-propagation, the adversarial training FGM method [4]is used to improve the performance and robustness of the model. (2) Multiple pre-trained language models ensemble. We have trained several models that perform well in the field of ABSA for enhancing the understanding of the DiaASQ task. Given that different models may learn different dialog thread features, we adopts a voting mechanism and ensemble learning to improve the performance of the CASQA system [5]. We perform ensemble learning on the 5 different pre-trained language models to obtain corresponding model combinations, and the final triplets prediction results is obtained by internal voting among the models. During the ensemble learning process, we propose a model ensemble algorithm called beam search ensemble.

In summary, our contributions are as follows:

- We use the optimized RoPE to further improve the model's understanding of dialog context and adversarial training to improve the robustness of the model.
- To leverage the distinctive dialog thread features learned by different pre-trained models, we employ the beam search ensemble algorithm. This algorithm merges the predicted results from these models, allowing us to integrate their insights and enhance the overall performance.
- Our proposed system achieved the second place in the Chinese dataset and fifth place in the general dataset during the final evaluation of the Conversational Aspect-based Sentiment Quadruple Analysis shared task at NLPCC-2023. This achievement strongly demonstrates the effectiveness of our method.

2 Related Work

From the traditional approach of text-level sentiment analysis to the more comprehensive fine-grained analysis that encompasses opinion mining through the prediction of various elements, including aspect terms, sentiment polarity, opinion terms, aspect categories, and targets. The growing popularity of open-domain

dialogue systems, particularly ChatGPT, has given rise to increased interest in sentiment analysis of integrated conversations.

Zhao et al. [6] and Wu et al. [7] proposed an end-to-end method to solve the task of Pair-wise Aspect and Opinion Terms Extraction and a multi-task learning framework based on shared spans, where the terms are extracted under the supervision of span boundaries. Peng et al. [8] proposed a two-stage framework to extract aspect sentiment triplet. The first stage predicts what, how and why in a unified model, and then the second stage pairs up the predicted what (how) and why from the first stage to output triplets. Knoester et al. [9] proposed work extends a state-of-the-art model for ABSA with the methodology of Domain Adversarial Training to create a deep learning adaptable cross-domain structure. This improves the generalization and robustness of the model. Li et al. [2] constructed a large-scale high-quality DiaASQ dataset which contains both Chinese and English version. They bridged the gap between fine-grained sentiment analysis and conversational opinion mining by developing a neural model which shows huge wins on the cross-utterance quadruple extraction. However, their systems have limited understanding of the entity, aspect, and sentiment triples in multi-turn dialogues. In contrast, our optimized rotational position embedding enables our model to better comprehend the relationships between triples across the conversation context. Additionally, our proposed model integration method leverages multiple perspectives to enhance the accuracy of triplet extraction in the model's multi-turn conversation flow.

3 Methodology

3.1 Triplets Extraction Model

Based on tree-structured parzen estimator (TPE), a classic Bayesian optimization algorithm, we obtain a preliminary range of hyperparameters suitable for different models. On various long text benchmark datasets, Su et al. [3] proposed Rotary Position Embedding(RoPE). By using RoPE, various valuable properties can be achieved, such as the ability to flexibly adjust sequence length, a reduction in the strength of inter-token dependencies at greater relative distances, and the potential to enhance the linear self-attention mechanism with relative position encoding. Consistent superior performance in comparison to alternative methods has been demonstrated through their experiences. Our task is to integrate the dialog into the tree-like dialogue replying structure. Since our context length is shorter than the long text dataset on Roformer, we modify the weight of the rotation positional information embedding.

$$\begin{cases} \boldsymbol{p}_{i,2t} & = \sin\left(10000^{-wei*2t/d}\right) \\ \boldsymbol{p}_{i,2t+1} & = \cos\left(10000^{-wei*2t/d}\right) \end{cases} \tag{1}$$

in which wei is the RoPE embedding weight that we adjusted and $p_{i,2t}$ is the $2t^{th}$ element of the d-dimensional vector p_i.

FGM is an adversarial training method, applying adversarial perturbations to word embeddings. Suppose the word embedding vector is $\boldsymbol{s}$, and the model conditional probability of y given $\boldsymbol{s}$ as $p(y|s;\theta)$, where θ are the parameter of the classifier, $\boldsymbol{N}$ is the number of labeled examples. Then the adversarial perturbation $\boldsymbol{r_{adv}}$ on $\boldsymbol{s}$ as

$$\boldsymbol{r}_{\text{adv}} = -\boldsymbol{\epsilon g}/\|\boldsymbol{g}\|_2 \text{where} \boldsymbol{g} = \nabla_s \log p(y \mid \boldsymbol{s}; \boldsymbol{\theta}). \tag{2}$$

The adversarial loss is computed as

$$\boldsymbol{L}_{\text{adv}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N} \log p(y_n \mid s_n + \boldsymbol{r}_{\text{adv},n}; \boldsymbol{\theta}) \tag{3}$$

Based on the above optimization strategy, we add 100 dialogue verification sets to the training set to train the model, and finally get the best performance of a single model. The structure of our system is shown in Fig. 2.

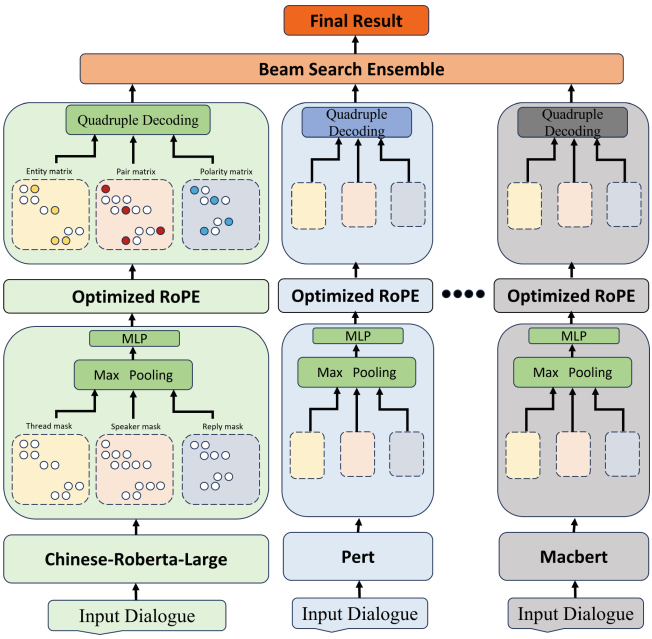

Fig. 2. The overall architecture of the system

3.2 Models Selection

In this section, we conducted experiments using several powerful Aspect-Based Sentiment Analysis (ABSA) systems that have been verified on classic ABSA

tasks, as reported in [10]. To encode the dialogue text for our task, we trained both the English and Chinese versions of these models separately.

BERT [11] was the first pre-trained language model that used a large-scale corpus, and has led to significant performance improvements in many downstream natural language processing tasks. In recent years, several improved Chinese pre-trained language models based on BERT have emerged, including the Chinese versions of RoBERTa-wwm [12], PERT [13], and MacBERT [14].

RoBERTa-wwm is a Chinese pre-trained language model based on RoBERTa that uses a whole-word masking strategy and other pre-training techniques to improve performance.

PERT takes a different approach to pre-training by replacing the Masked Language Model (MLM) with a word order prediction task, where the model is presented with randomly shuffled text and tasked with predicting the original word order. This approach has been shown to improve the performance of pre-trained models.

MacBERT improves upon the pre-training technique of RoBERTa by incorporating a synonym masking strategy. This strategy aims to reduce the gap between pre-training and fine-tuning phases, and has demonstrated effective performance improvements in Chinese pre-trained language models.

These models were selected for their proven efficacy in ABSA tasks and were separately trained for Chinese and English language inputs.

3.3 Ensemble Model

As the number of distinct models continues to increase, finding the optimal combination of models quickly becomes computationally expensive to model with traditional methods. To overcome this challenge, we propose a beam search ensemble algorithm for model fusion. This algorithm incorporates improvements to the beam search approach, enabling a more efficient convergence to the optimal combination. As shown in Fig. 3, we use beam search ensemble algorithm to get the final result.

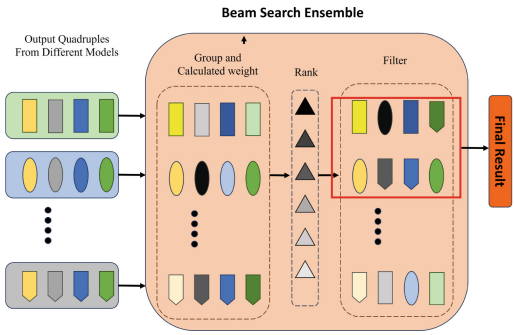

Fig. 3. The details of the beam search ensemble

One key issue with traditional model fusion is the distribution of voting weight amongst the models. In our algorithm, we have effectively addressed this issue by considering the performance of each model and reducing any bias that may arise from poorly performing models. By doing so, the algorithm ensures that only models with good performance are given more significant voting weight, and thus produce optimal results.

Our proposed algorithm is superior in terms of generalization ability when compared to other existing model fusion methods. In addition, the algorithm reduces time complexity significantly and allows for single model voting. This approach fully considers the strengths of each model, and effectively updates the voting weight to produce the best possible results. Overall, the beam search ensemble algorithm can combine the prediction results of all models to obtain more accurate prediction results. As follows.

Beam search ensemble algorithm to merge distinct predicted results

Input: models_data: a list of multiple model predictions

output: final_data

set **k**: the most likely number of emotions to be selected in each identifier combination

set threshold **t** ,define empty dictionary: ensembled_data, define empty lists: final_data, beam

```
for data in models_data do
    for doc in data do
        for triplet in doc['triplets'] do
            identifier=(doc['doc_id'],)+tuple(triplet[:-3])
            ensembled_data[identifier].append(triplet[-3])
        end for
    end for
end for
for identifier, predictions in ensembled_data.items() do
    counter = a statistical emotion counter
    for prediction in predictions do
        counter[prediction] += 1
    end for
    if the counter length >= 1 do
    beam_sum =the sum of emotional counts in beam
    emotion_weight = count / beam_sum
        if emotion_weight >= threshold t do
        add quadruple to final_data
        end if
    end if
end for
```

In this process, the calculation of the importance weight is to calculate the relative weight of each emotion in proportion through the number of occurrences of each emotion in the prediction result of the statistical model, and then filter the prediction result according to the threshold. Compared with the voting method, the beam search ensemble algorithm can better deal with the situation where there are fewer identifiers and the emotional distribution is unbalanced, and it can better control the misjudgment rate while improving the prediction effect.

4 Experiments

4.1 Dataset and Evaluation Index

The DiaASQ dataset includes 1000 dialogues in both Chinese and English languages. It is split into a training set, a verification set, and a test set at a ratio of 8:1:1 for each language. Since the data was originally in Chinese, and the English data set was translated from it, there is some degree of noise in the English data, which accounts for the lower F1 scores of the model on the English data set compared to the Chinese data set. As a result, improving the model's performance on the English data set proves to be more challenging. Since our main focus is on quadruple extraction, we primarily measure the performance using micro F1 and identification F1 scores. The micro F1 score considers the entire quad, including the sentiment polarity. On the other hand, the identification F1 score, as described in Barnes et al. [15], does not differentiate between different polarities in the evaluation.

4.2 Results and Analysis

To evaluate the effectiveness of the optimization techniques we applied to our models, we first selected RoBERTa-large, the most effective pre-trained language model, and conducted a single-model comparison experiment on both the Chinese and English test sets. Next, we compared each baseline model to the model incorporating all of our improvements. In the end, we will employ the beam search ensemble algorithm to obtain the best possible prediction result by leveraging the various models optimized for optimal performance. +FGM means that adversarial training modules are added to the +optimized RoPE, +verification sets means that model training is further trained with verification sets on the basis of the first two and +all means using all of the above strategies at the same time. The ablation experiments are shown in Table 2.

Table 2 demonstrates that incorporating optimized RoPE into the DiaASQ baseline results in an improvement of 1.02% and 0.87% for the roberta-large model in both Chinese and English datasets. This highlights the effectiveness of adjusting the weight of the positional embedding information based on the length of the conversation and its ability to enhance the model's context comprehension. Additionally, the adversarial training process that accumulates both the original

Table 2. The F1 of different models on the Chinese and English test set

Model	Chinese			English		
	Micro-F1	Iden-F1	Avg. F1	Micro-F1	Iden-F1	Avg. F1
DiaASQ (baseline)	34.94	37.51	36.23	33.31	36.8	35.06
roberta-large	36.24	42.78	38.31	33.67	37.3	34.69
+optimized RoPE	38.35	42.72	39.33	34.74	37.96	35.55
+FGM	38.50	42.92	39.51	34.82	38.08	35.65
+verification sets	39.23	43.85	**41.54**$_{+5.32}$	35.13	38.55	**36.84**$_{+1.79}$
roberta-base	33.7	40.02	36.86	31.68	35.96	33.82
+all	35.74	44.26	**40.00**	32.74	37.2	**34.97**
bert-large	33.17	39.33	36.25	31.29	35.41	33.35
+all	35.53	43.9	**39.715**	32.67	36.98	**34.83**
Pert	31.28	38.66	34.37	/	/	/
+all	34.64	42.83	**38.135**	/	/	/
Macbert	33.05	40.79	36.92	/	/	/
+all	35.39	43.85	**39.62**	/	/	/
beam search ensemble(ours)	42.09	45.62	**43.855**$_{+7.63}$	35.54	38.76	**37.15**$_{+2.10}$

gradient and the adversarial gradient can mitigate overfitting and improve the model's robustness. After including a verification set consisting of 100 dialogues in the training data, the average F1 score for the Chinese dataset increased by 2.03% in the Roberta-large model, while that of the English dataset increased by 1.21%. This suggests that the dataset is of high quality, and that the amount of data plays a significant role in limiting the performance of the models. We applied the aforementioned optimization techniques to other Bert-based models, and observed improvement in their performance. Considering the different features learned by each model and the nuances of their predicted quadruples, we utilized the beam search ensemble algorithm to merge the predictions of multiple models in the Table 2, assign weights to each of them, sort them, and screen out the quadruples with weights greater than the threshold t. The final result showed an increase of 7.63% compared to the baseline. Furthermore, when compared to the optimal model, the beam search ensemble algorithm demonstrated an improvement of 2.31%.

5 Conclusion

In this paper, we propose a model ensemble approach for conversational quadruple extraction. We initiated our efforts to enhance the task's performance by optimizing the RoPE positional information embedding. Subsequently, we employed adversarial training techniques to further boost the model's robustness and generalization capabilities. Additionally, we expanded the training dataset and further improved the model's F1 scores for both Chinese and English test set. We then trained several models using these optimization strategies and identified

the best results using the beam search ensemble algorithm. Experimental results on the NLPCC2023 Shared Task4 DiaASQ dataset demonstrate the effectiveness of our method and the necessity of the rotation positional information embedding module and using beam search ensemble algorithm to integrates correct predictions from distinct models.

References

1. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining Text Data, pp. 415–463. Springer (2012)
2. Li, B., Fei, H.: Diaasq: a benchmark of conversational aspect-based sentiment quadruple analysis. CoRR abs/2211.05705 (2022)
3. Su, J., Lu, Y.: Roformer: enhanced transformer with rotary position embedding. CoRR abs/2104.09864 (2021)
4. Miyato, T., Dai, A.M.: Adversarial training methods for semi-supervised text classification. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net (2017)
5. Cui, S., Han, Y.: A two-stage voting-boosting technique for ensemble learning in social network sentiment classification. Entropy **25**(4), 555 (2023)
6. Zhao, H., Huang, L.: Spanmlt: a span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, pp. 3239–3248 (2022)
7. Wu, S., Fei, H.: Learn from syntax: improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In: Zhou, Z. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19–27 August 2021, pp. 3957–3963 (2021)
8. Peng, H., Xu, L.: Knowing what, how and why: a near complete solution for aspect based sentiment analysis. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, pp. 8600–8607. AAAI Press (2020)
9. Knoester, J., Frasincar, F.: Domain adversarial training for aspect-based sentiment analysis. In: Web Information Systems Engineering - WISE 2022–23rd International Conference, Biarritz, France, November 1–3, 2022, Proceedings. LNCS, vol. 13724, pp. 21–37 (2022)
10. Li, Z., Zou, Y.: Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event/Punta Cana, Dominican Republic, 7–11 November, 2021, pp. 246–256. Association for Computational Linguistics (2021)
11. Devlin, J., Chang, M.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)

12. Liu, Y., Ott, M.: Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019)
13. Cui, Y., Yang, Z.: PERT: pre-training BERT with permuted language model. CoRR abs/2203.06906 (2022)
14. Cui, Y., Che, W.: Revisiting pre-trained models for Chinese natural language processing. In: Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020. Findings of ACL, vol. EMNLP 2020, pp. 657–668
15. Barnes, J., Kurtz, R.: Structured sentiment analysis as dependency graph parsing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1–6, 2021