Adversarial Training and Model Ensemble for User Feedback Prediciton in Conversation System

Junlong Wang[1], Yongqi Leng[2], Xinyu Zhai[1], Linlin Zong[1], Hongfei Lin[3], and Bo Xu[3(✉)]

[1] School of Software Technology, Dalian University of Technology, Dalian, China

[2] College of Intelligence and Computing, Tianjin University, Tianjin, China
[3] School of Computer Science and Technology, Dalian University of Technology, Dalian, China
xubo@dlut.edu.cn

Abstract. Developing automatic evaluation methods that are highly correlated with human assessment is crucial in the advancement of dialogue systems. User feedback in conversation system provides a signal that represents user preferences and response quality. The user feedback prediction (UFP) task aims to predict the probabilities of likes with machine-generated responses given a user query, offering a unique perspective to facilitate dialogue evaluation. In this paper, we propose a powerful UFP system, which leverages Chinese pre-trained language models (PLMs) to understand the user queries and system replies. To improve the robustness and generalization ability of our model, we also introduce adversarial training for PLMs and design a local and global model ensemble strategy. Our system ranks first in NLPCC 2023 shared Task 9 Track 1 (User Feedback Prediction). The experimental results show the effectiveness of the method applied in our system.

Keywords: User Feedback Prediction · Pre-trained Language Model · Adversarial Training · Model Ensemble

1 Introduction

Open-domain conversational system is designed to satisfy users' need [1] such as information support, communication, and entertainment, etc. Appraising the quality of the responses not only reflects the system's capability but also offers valuable insights for identifying areas that require further improvements [2]. In order to get the user's explicit satisfaction with the generated responses by machine, online conversation systems usually have a user feedback mechanism, such as like and dislike buttons, users can click the like button when they are satisfy with the response, and vice versa for the dislike button. These real-world feedback signal represents the user's vote on the quality of the response and also

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 366–375, 2023.
https://doi.org/10.1007/978-3-031-44699-3_33

represents their preference. Based on this, we can study how to cater to user's preferences and improve the quality of the generated responses to obtain high likes.

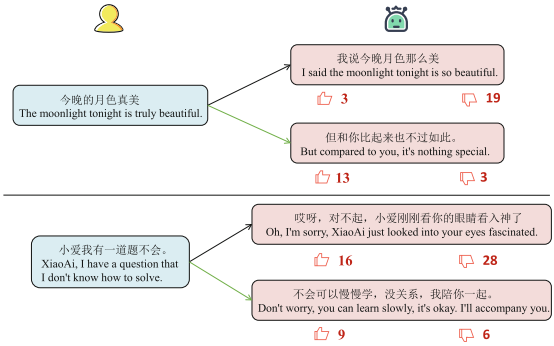

Fig. 1. Two examples from the evaluation task dataset.

In this paper, we focus on solving the problem of UFP in conversation system. Figure 1 present two examples from the evaluation task dataset, in the first example located in the upper half, we can see that the reply that praises the user and is more human-like receives positive feedback with high likes, while the reply that receives negative feedback with high dislikes appears repetitive and boring. In this task, we aim to predict the probabilities of likes given a query-reply pair (e.g., 3/22 and 13/16 for the first example). In order to better understand user preferences for response content in dialogue scenarios and objectively measure the quality of system responses, we conducted a study based on the real-world user feedback dataset from the dialogue system in NLPCC2023 shared task. In this work, we propose an effective UFP system that leverages a local and global ensemble of multiple Chinese PLMs, incorporating adversarial training for fine-tuning.

In the preprocessing phase, we perform purposeful data augmentation to expand the training data. Subsequently, we formulate the UFP task as a text regression problem and employ PLMs with Muti-Sample Dropout (MSD) [3] to locally approximate the ensemble of multiple models. Additionally, we introduce adversarial training [4] to substantially enhance the robustness of individual models.

In the end, we significantly improved the performance of our system by employing blending ensemble strategy, which integrates multiple independent models globally. The result shows that the Kullback-Leibler similarity score of the system proposed in this paper for UFP is 92.13, ranking the first, which indicates that the effectiveness and superiority of our system. The main contributions of this paper can be summarized as follows:

- We build our system based on the Chinese PLMs to learn the user satisfaction level for user queries and system replies in UFP task.
- We integrate adversarial training to effectively enhance the robustness of individual models and devise a local and global ensemble method, which significantly improves the performance of the UFP system.
- The final evaluation results show that our proposed system achieves the first place in the contest, which proves the effectiveness of our method.

2 Related Work

Due to the openness of content and the diversity of topics, it is challenging to evaluate the quality of responses in open domain dialogue systems. Existing automatic evaluation methods usually assess the response from language quality (Perplexity [5]), response diversity (DIST [6]), and relevance (BLEU [7]), they are easy to conduct but ineffective to reflect the dialogue quality [8]. Relatively, human evaluation is of high reliability but it tends to be very cost- and time-intensive. Therefore, researchers have made great efforts to find more reliable automatic evaluation methods that are highly correlated with human evaluation. USR [9] leveraged an unsupervised and reference-free method to approximate the specific scores rated by annotators. SelF-Eval [10] designed a self-supervised fine-grained dialogue evaluation model with a multilevel contrastive learning method. Sun et al. [11] proposed a user satisfaction annotation dataset for dialogue evaluation.

BERT-based [12] PLMs have facilitated the understanding of the relationship between context and response in dialogue evaluation tasks [13], and many improved Chinese PLMs from BERT have emerged in recent years. RoBERTa-wwm [14] use whole word masking strategy based on RoBERTa [15] . MacBERT [16] achieves significant results in Chinese NLP tasks through adopting masked language model as correction. ERNIE [17] strengthens representations by masking knowledge entities from the corpus. Moreover, in order to improve the performance of specific downstream tasks, many works have introduced adversarial training [18,19] for model to improve robustness to small, approximately worst case perturbations. Additionally, in many shared tasks, the winners integrate multiple models instead of using a single model [20,21]. Ensemble learning methods leverage multiple models to extract different features from the training data and then combine the prediction results through various strategies such as Bagging, Boosting, Stacking, Voting [22], and Blending, etc. Ensemble learning effectively reduces errors in individual models, and improves generalization ability.

3 Methodology

Our method is divided into two stages: (1) single text regression model training, where each query-reply pair in the training data is concatenated as input to the

Encoder to extract semantic information. Then, we apply MSD for local model integration, and a regression head is used for prediction. After gradient back-propagation, PGD adversarial training is employed to enhance the robustness of the model. Additionally, targeted data augmentation is applied based on the characteristics of the dataset. The architecture of our model is shown in Fig. 2. (2) multi-model ensemble, we globally integrate the models obtained from the first stage using the blending ensemble method.

3.1 Task Definition

Given a user query and system reply pair, our goal is to predict the probabilities of likes p_i for it, where $p_i \in [0, 1]$, and the computation of p_i is determined by the relative frequency of the term "like" within user feedback. Obviously, we can define UFP task as a regression problem.

3.2 Model Architecture

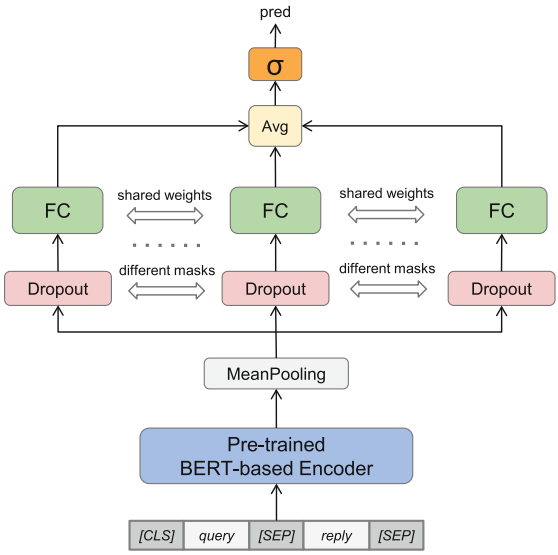

Fig. 2. The architecture of our model

Input Representation. For each query-reply pair, we concatenate the query and reply sentences into a text sequence in the following form:

$$\{[CLS], query, [SEP], reply, [SEP]\}$$

We utilize a BERT-based encoder to represent the input sequence. Subsequently, we apply the mean pooling to the output of the Encoder to obtain the sentence-level embedding.

Muti-Sample Dropout for Local Ensemble. Before the regression head, which consists of a fully connected layer and a sigmoid activation function, we employ MSD [3] as a regularization technique. Specifically, we incorporate multiple parallel dropout layers and they share the fully connected layer, then we can compute the average of the logits from these parallel samples. This module approximates the ensemble of multiple sub-models by averaging the predicted results obtained from different inputs, effectively enhancing the generalization ability of the model.

3.3 Adversarial Training

We introduce Projected Gradient Descent (PGD) [23], which is considered to be the best in first-order adversarial, to improve the robustness of the model in UFP task. PGD adopts a multi-iteration approach to obtain the optimal adversarial examples within a batch, in contrast to Fast Gradient Descent [18]. To be specific, the word embedding layer of the PLMs is attacked in each iteration. Finally, the adversarial examples generated by multiple iterations are superimposed to obtain the optimal value. The adversarial examples generation formula is as follows.:

$$e_{t+1} = \prod_{e+S}(e_t + \alpha g(e_t)/||g(e_t)||_2) \tag{1}$$

$$g(e_t) = \nabla_e L(e_t, y) \tag{2}$$

where e_t is adversarial examples generated at step t, $g(\cdot)$ represents the gradient calculation function. Both α and S are hyperparameters representing step size and adversarial perturbation space, respectively.

3.4 Data Augmentation

In Fig. 3, we can observe the presence of imbalanced distribution in the dataset, and as show in Table 3, the average number of likes per reply in the training set is twice that of dislikes, this disparity is even more pronounced in the test set. Intuitively, we can improve the model's ability to extract the features of replies that users prefer and mitigate label imbalance issues by expanding high-like data. To this end, we have devised two targeted data augmentation strategies:

1-R: we only augment the replies of the *All* data in Fig. 3.

1-QR: based on **1-R**, we also expand the queries of the samples.

Specifically, for a query or a reply sentence, we randomly select one of four strategies to generate a pseudo sentence: *Back translation*[1], *Replacement with synonyms*, *Deletion words*, and *Exchange adjacent characters*.

[1] We implemented it by calling the Baidu Translation API:https://api.fanyi.baidu.com/api.

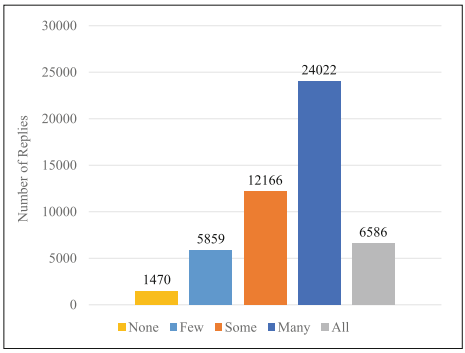

Fig. 3. The label distribution of the training set. *None* represents replies where all user feedbacks are dislike ($p_i = 0$), *Few* denotes $p_i \in (0, 0.3]$, *Some* denotes $p_i \in (0.3, 0.6)$, *Many* represents that $p_i \in [0.6, 1)$, and *All* represents replies where all user feedbacks are like ($p_i = 1$).

3.5 Blending for Global Ensemble

In order to integrate the learning ability of each model and improve the generalization ability of the final system, we use a blending ensemble strategy to integrate multiple models. For each model, we select the strategy that yield the best performance and utilize the predictions on the validation set by all fine-tuned models as features for the second stage, where we learn a new model to ensemble them. We employ linear regression method to learn the weights for model fusion, the final system is obtained by weighting the contributions of five individual models.

4 Experiments

As shown in Table 1, our system achieved the 1st place in the final official test set results. We conducted a three-stage experiment to prove the effectiveness of our method. First, we selected five effective PLMs to conduct a single-model comparison experiment on the validation set as baselines: RoBERTa[2] [16], ERNIE-3.0-xbase[3] [17], ERNIE-3.0-base[4] [17], MacBERT[5] [16], and MRC-RoBERTa[6]. We used the root mean square error loss and some main training hyper-parameters are shown in Table 2. Secondly, we conducted comparative experiments for each baseline model that incorporated all improvements, the results are show in Table 4. Finally, the model ensemble experiment was conducted.

[2] https://huggingface.co/hfl/chinese-roberta-wwm-ext-large.
[3] https://huggingface.co/nghuyong/ernie-3.0-xbase-zh.
[4] https://huggingface.co/nghuyong/ernie-3.0-base-zh.
[5] https://huggingface.co/hfl/chinese-macbert-large.
[6] https://huggingface.co/luhua/chinese_pretrain_mrc_roberta_wwm_ext_large.

Table 1. Final results of top-5 teams.

System name	KL-Score	Rank
Ours	**92.13**	**1**
dunnlp	92.00	2
zut	91.73	3
YNU-HPCC	91.63	4
HTDZNLP	91.40	5

Table 2. Hyper-parameter setting.

Setting	Value
Epoch num	8
Batch size	32
Optimizer	Adam
Learning rate	1e-5
Dropout num of MSD	5
Dropout rate of MSD	0.3

4.1 Dataset and Evaluation

The overview statistics of dataset is shown in Table 3. A query sentence may have multiple replies, and we consider a query-reply pair as a single sample.

Table 3. Statistics of the evaluation task dataset.

Item	train	val	test
# Query	16000	2000	2000
Avg # Reply	3.14	3.07	3.16
Avg # Like per Reply	16.15	19.84	30.57
Avg # Dislike per Reply	8.42	9.41	12.19

In this task, the evaluation metric is Kullback-Leible similarity score (KL-Score), for a gold and a prediction, the score is computed as:

$$\begin{aligned} Score1 &= gold \times log(gold) + (1 - gold) \times log(1 - gold) \\ Score2 &= gold \times log(pred) + (1 - gold) \times log(1 - pred) \\ KL\text{-}Score &= \frac{1}{1 + Score1 - Score2} \end{aligned} \tag{3}$$

then, the average score of all replies for each query is calculated, and the final result is the average score across all queries.

4.2 Results and Analysis

(1) Among all the base models, ERNIE-xbase performs the best, while the smaller parameter-sized ERNIE-base also shows impressive performance. Moreover, ERNIE-xbase+PGD+1-QR achieves the best KL-Score of 91.590, demonstrating its powerful Chinese semantic understanding ability through knowledge enhanced pre-training methods.

Table 4. The KL-Score of different models on the validation set.

Method	RoBERTa	ERNIE-xbase	ERNIE-base	MacBERT	MRC-RoBERTa
Base	91.263	91.353	91.272	91.220	91.207
+MSD	91.289	91.353	91.281	91.053	91.420
+MSD+PGD	**91.485**	91.500	91.336	91.356	**91.584**
+MSD+PGD+1-R	85.469	91.583	**91.532**	91.356	91.384
+MSD+PGD+1-QR	94.363	**91.590**	91.503	**91.365**	91.374

Table 5. The KL-Score of model ensemble on the validation and test set.

Models	Val	Test
ERNIE-xbase+RoBERTa	91.730	92.038
ERNIE-xbase+RoBERTa+MacBERT	91.768	92.065
ERNIE-xbase+RoBERTa+MacBERT+ERNIE-base	91.802	92.127
ERNIE-xbase+RoBERTa+MacBERT+ERNIE-base+MRC-RoBERTa	**91.818**	**92.136**

(2) The MSD local ensemble method improves the performance of all models except MacBERT. Additionally, the experimental results highlight the effectiveness of PGD adversarial training, all models have shown remarkable performance improvement after the introduction of PGD, particularly with RoBERTa, which achieved a 0.196 improvement in terms of the KL-Score. MRC-RoBERTa with MSD and PGD achieves the second-best overall performance. This model is based on RoBERT-wwm pre-trained on a large-scale reading comprehension corpus, and we speculate that the UFP task can also be viewed as a reading comprehension task, where its strong reading comprehension ability effectively models the matching degree between queries and replies.

(3) We also experimented with two data augmentation strategies on each model, the results show that ERNIE-xbase and ERNIE-base achieved the best performance through the 1-QR and 1-R strategies, respectively, validating the effectiveness of our purposeful augmentation of high-like data.

(4) In terms of model ensemble, as shown in Table 5, the blending method for global integration significantly improves the performance of the UFP system. Furthermore, the system's performance is directly proportional to the number of models in the ensemble, demonstrating the power and scalability of the ensemble method.

5 Conclusion and Future Work

In this work, we define the UFP task as a text regression task that measures the quality of responses and understands user preferences in conversation system. We perform purposeful data augmentation and utilize PLMs to encode queries and replies, and introduce adversarial training and a local and global model

ensemble strategy to significantly improve the system's robustness and performance. Experimental results show the effectiveness of our proposed method, and our system achieved first place in track 1 of the NLPCC 2023 Shared Task 9. The effective UFP system demonstrates its applicability to dialogue evaluation tasks. Moreover, it can be further extended to encompass various query-reply matching tasks, facilitating the retrieval or generation of highly scored replies to meet user requirements.

In the future, we will focus on mining the differences between replies with different scores to the same query, to efficiently utilize user feedback as a supervisory signal for designing models that are better suited for dialogue evaluation tasks.

Acknowledgements.. This work was supported in part by the National Natural Science Foundation of China under Grant Grant 62006034; in part by the Natural Science Foundation of Liaoning Province under Grant 2021-BS-067; and in part by the Dalian High-level Talent Innovation Support Plan under Grant 2021RQ056.

References

1. Shum, H., He, X., Li, D.: From Eliza to XiaoIce: challenges and opportunities with social chatbots. Front. Inf. Technol. Electron. Eng. **19**(1), 10–26 (2018). https://doi.org/10.1631/FITEE.1700826
2. Deriu, J., et al.: Survey on evaluation methods for dialogue systems. Artif. Intell. Rev. **54**(1), 755–810 (2021). https://doi.org/10.1007/s10462-020-09866-x
3. Inoue, H.: Multi-sample dropout for accelerated training and better generalization. arXiv preprint arXiv:1905.09788 (2019)
4. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
5. Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. In: Advances in Neural Information Processing Systems 13 (2000)
6. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. (2002)
8. Liu, C.-W., Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023 (2016)
9. Mehri, S., Eskenazi, M.: USR: an unsupervised and reference free evaluation metric for dialog generation. arXiv preprint arXiv:2005.00456 (2020)
10. Ma, L., Zhuang, Z., Zhang, W., Li, M., Liu, T.: Self-Eval: self-supervised fine-grained dialogue evaluation. arXiv preprint arXiv:2208.08094 (2022)
11. Sun, W., et al.: Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2499–2506 (2021)
12. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

13. Ye, Z., Lu, L., Huang, L., Lin, L., Liang, X.: Towards quantifiable dialogue coherence evaluation. arXiv preprint arXiv:2106.00507 (2021)
14. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans. Audio, Speech, Lang. Process. **29**, 3504–3514 (2021)
15. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
16. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922 (2020)
17. Sun, Y., et al.: ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137 (2021)
18. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725 (2016)
19. Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Zhao, T.: SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. arXiv preprint arXiv:1911.03437 (2019)
20. Agrawal, S., Mamidi, R.: Lastresort at semeval-2022 task 4: towards patronizing and condescending language detection using pre-trained transformer based models ensembles. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 352–356 (2022)
21. Yu, W., Boenninghoff, B., Roehrig, J., Kolossa, D.: Rubcsg at semeval-2022 task 5: ensemble learning for identifying misogynous memes. arXiv preprint arXiv:2204.03953 (2022)
22. Sagi, O., Rokach, L.: Ensemble learning: a survey. Wiley Interdisc. Rev.: Data Min. Knowl. Discov. **8**(4), e1249 (2018)
23. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)