Retrieve, Discriminate and Rewrite: A Simple and Effective Framework for Obtaining Affective Response in Retrieval-Based Chatbots

Xin Lu, Yijian Tian, Yanyan Zhao, Bing Qin*
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{xlu, yjtian, yyzhao, qinb}@ir.hit.edu.cn

Abstract

Obtaining affective response is a key step in building empathetic dialogue systems. This task has been studied a lot in generation-based chatbots, but the related research in retrieval-based chatbots is still in the early stage. Existing works in retrieval-based chatbots are based on Retrieve-and-Rerank framework, which have a common problem of satisfying affect label at the expense of response quality. To address this problem, we propose a simple and effective Retrieve-Discriminate-Rewrite framework. The framework replaces the reranking mechanism with a new discriminate-and-rewrite mechanism, which predicts the affect label of the retrieved high-quality response via discrimination module and further rewrites the affect unsatisfied response via rewriting module. This can not only guarantee the quality of the response, but also satisfy the given affect label. In addition, another challenge for this line of research is the lack of an off-the-shelf affective response dataset. To address this problem and test our proposed framework, we annotate a Sentimental Douban Conversation Corpus based on the original Douban Conversation Corpus. Experimental results show that our proposed framework is effective and outperforms competitive baselines.

1 Introduction

Expressing affect is a key factor to build human-like dialogue systems, which can significantly promote affective communication and enhance user satisfaction (Prendinger and Ishizuka, 2005; Partala and Surakka, 2004) during human-computer interactions. This problem has been studied a lot in generation-based chatbots (Zhou et al., 2018; Zhou and Wang, 2018; Song et al., 2019; Shen and Feng, 2020), which is usually defined as obtaining an affective response given an affect label and the context of a conversation (Yuan et al., 2020).

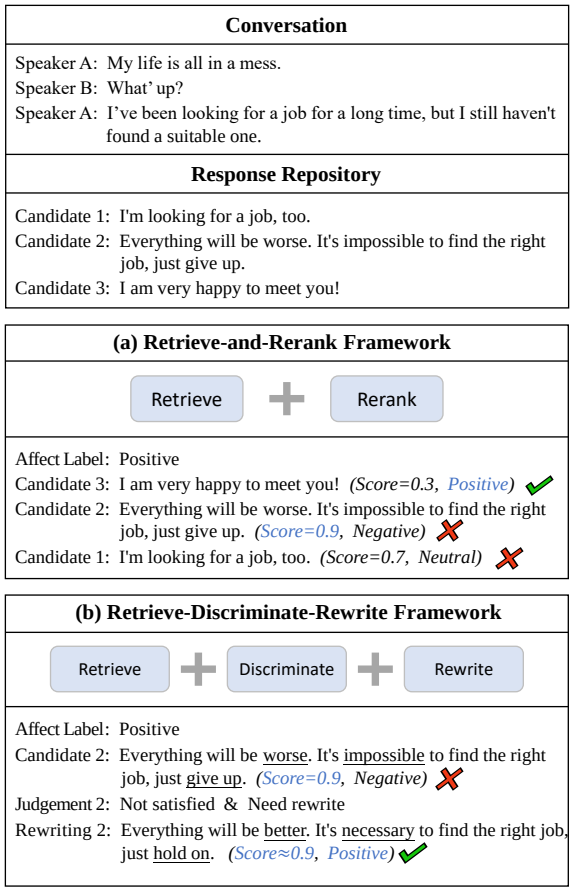


Figure 1: A short conversation example which shows the difference between two frameworks.

However, the related research in retrieval-based chatbots is still in the early stage (Qiu et al., 2020). Retrieval-based chatbots have an advantage over generation-based chatbots in obtaining diverse and informative responses, which is also widely used. Therefore, research on obtaining affective response in retrieval-based chatbots is meaningful. In existing studies, affect is regarded as the term that subsumes emotion, feeling and sentiment (Fleckenstein, 1991). In this paper, we focus on sentiment and study how to obtain a specific polarity (*positive* or *negative*) response in retrieval-based chatbots.

* Email corresponding.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1956–1969
November 7–11, 2021.

Different from generation-based chatbots that can generate new responses, retrieval-based chatbots must obtain responses based on the candidates retrieved from a response repository. Therefore, under the objective of obtaining affective response, how to effectively use the candidates is an important issue. Existing works in retrieval-based chatbots are based on Retrieve-and-Rerank framework (Lubis et al., 2019; Qiu et al., 2020), which employs a reranking mechanism to the retrieved candidates. Specifically, the framework firstly obtains the candidates via retrieval module, then adjusts ranking or matching score according to the given affect label, and finally outputs a response that is appropriate in both affect and content.

However, the Retrieve-and-Rerank framework is not sufficient since it satisfies given affect label at the expense of response quality (Qiu et al., 2020). This means that high-quality but affect unsatisfied responses will be discarded, which directly reduces the core advantage of retrieval-based chatbots. For example, in Figure 1(a), when affect label is not considered, high-quality candidate 2 should be the best one, but since the affect label is given, only candidate 3 with ordinary quality can be selected.

To guarantee content and affect at the same time, we propose a simple and effective Retrieve-Discriminate-Rewrite framework. The framework replaces the reranking mechanism with a new discriminate-and-rewrite mechanism, which preferentially selects high-quality candidate response and rewrites the response whose affect is discriminated to be unsatisfied. For example, in Figure 1(b), our new framework preferentially selects high-quality candidate 2, then discriminates that the affect of the candidate 2 is unsatisfied, and finally makes the affect of the candidate 2 satisfied with a small amount of modification. This shows that the framework can not only guarantee the quality of response, but also satisfy the given affect label.

In addition, another challenge for this line of research is the lack of an off-the-shelf affective dataset. Such a dataset can not only be used in our framework, but also necessary for existing methods which employs the reranking mechanism. To address this problem and test our framework, we annotate a Sentimental Douban Conversation Corpus based on the original Douban Conversation Corpus which is widely used by many previous works in retrieval-based chatbots. We conduct experiments on this dataset, and experimental results show that our framework with a simple architecture is effective and outperforms competitive baselines.

The contributions of this work are summarized as follows:

- We propose a Retrieve-Discriminate-Rewrite framework for obtaining affective response in retrieval-based chatbots, which solves the problem of low-quality responses in the Retrieve-and-Rerank framework.
- We annotate and publish an affective response dataset, which solves the problem of the lack of necessary dataset in this line of research.
- Experimental results on the dataset show that our framework is effective and outperforms competitive baselines.

2 Related Work

Existing works for obtaining affective response in dialogue systems can be categorized into two branches. The first category is the generation-based method, which generates affective response for given conversation context based on the Seq2Seq model (Shang et al., 2015; Sordoni et al., 2015). The generation-based method has the advantage of generating new responses and has been studied a lot (Zhou et al., 2018; Zhou and Wang, 2018; Song et al., 2019; Shen and Feng, 2020). The second category is the retrieval-based method, which obtains affective response for given conversation context based on the candidates retrieved from the response repository. The retrieval-based method has the advantage of obtaining diverse and informative responses (Song et al., 2018), which is still competitive compared to the generation-based method. This paper focuses on the second category.

Different from the generation-based method, the related research on the retrieval-based method for obtaining affective response is still in the early stage. Lubis et al. (2019) proposed a reranking strategy for positive emotion elicitation whose method can also be applied to obtain affective response. Qiu et al. (2020) presented an emotion-aware matching network, which incorporated emotional factors and realized emotional control. From the perspective of using candidates, these methods are all based on Retrieve-and-Rerank framework. Although these methods can already obtain affective responses, Qiu et al. (2020) observed that these methods prefer responses that satisfied given affect label, even if they are not relevant to the context,

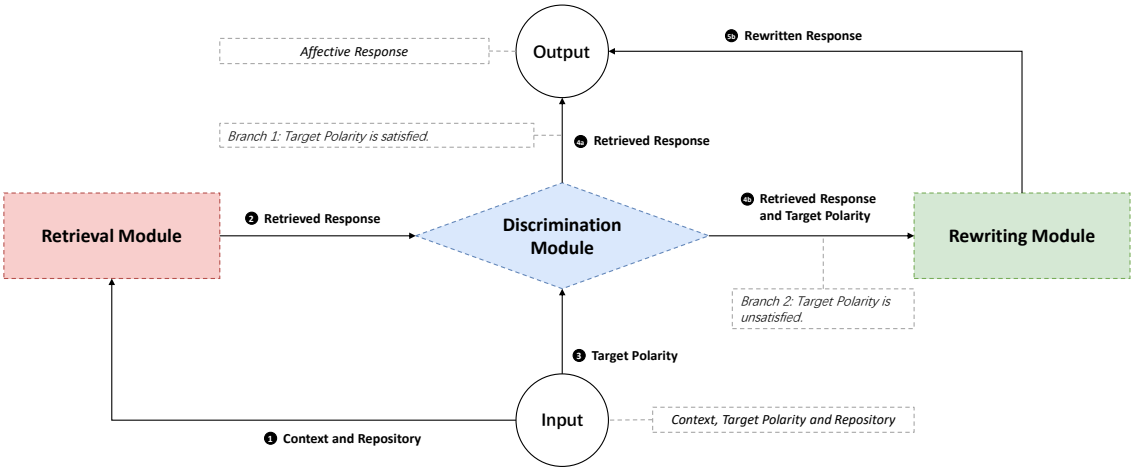


Figure 2: Overview of the Retrieve-Discriminate-Rewrite framework. We use digital numbers to show the processing flow of our framework. The framework includes three components: retrieval module, discrimination module and rewriting module. The retrieval module is used to retrieve a high-quality response, the discrimination module is used to discriminate the polarity of the response, and the rewriting module is used to correct the polarity of the response from unsatisfied to satisfied. These modules work together to obtain affective responses.

which reduces the core advantage of retrieval-based chatbots. How to balance rich information and given affect label is still being explored, and in this paper we focus on this problem.

Another branch of research touched in our work is style transfer in natural language processing. The rewriting mechanism in our framework will modify the polarity of the response, which has been studied in some style transfer works. Some existing works (Shen et al., 2017; Fu et al., 2018; Prabhumoye et al., 2018; Xu et al., 2018) focus on how to get style independent sentence representation and then generate a sentence with target style. These works have certain effectiveness, but they usually lack fine-grained control and cause poor content preservation, which is inconsistent with our goal of fine-grained control of polarities. Meanwhile, some existing works (Li et al., 2018; Sudhakar et al., 2019) focus on how to remove style-related words in the sentence and then generate a sentence with target style. Inspired by these works, we also regard the polarity rewriting as a similar two-stage process. But different from these works, our polarity rewriting will involve a large number of situations from neutral expression to affective expression, not just the transfer between different affective expressions. Lack of processing neutral expression will lead to poor performance in our task, and our handling of this problem is different from these works. In addition, some other existing works (Zhang et al., 2018; Lample et al., 2019; Dai et al., 2019) have realized style transfer from other different perspectives. These works aim at more general style transfer issues and also lack fine-grained control of polarities, which does not match the goal of our work.

3 The Retrieve-Discriminate-Rewrite Framework

3.1 Overview

In this work, our goal is to obtain an affective response given an affect label and the context of a dialogue in retrieval-based chatbots. In particular, the affect label we focus on is sentiment polarity (*positive* or *negative*).

The problem can be formulated as follows: given a conversation context $C = \{u_1, u_2, ..., u_N\}$ with N utterances, a response repository $P = \{r_1, r_2, ..., r_M\}$ related to the context C, and a target polarity s, the objective is to obtain a response Y based on the candidates retrieved from the response repository P, which not only is coherent with the context C, but also matches the target polarity s.

For this problem, our Retrieve-Discriminate-Rewrite framework is shown in Figure 2. The framework consists of three components:

(1) **Retrieval Module**: This module is used to be compatible with existing retrieval-based chatbots, which can provide high-quality response for

subsequent modules.

(2) **Discrimination Module**: This module receives the retrieved high-quality response from the retrieval module, which can discriminate the polarity of the retrieved high-quality response and output the response with satisfied polarity.

(3) **Rewriting Module**: This module receives the response with unsatisfied polarity from the discrimination module, which can correct the polarity of the response from unsatisfied to satisfied.

In the following sections, we will describe these components in detail, and introduce how the framework uses them to obtain affective response.

3.2 Retrieval Module

In our framework, the retrieval module is used to be compatible with existing retrieval-based methods. To verify that our framework is universal, we select the following retrieval-based methods to obtain high-quality responses in our framework, and we will conduct experiments based on these methods separately.

GTM This is the Ground Truth model, which always outputs correct responses. We use this ideal model to study the performance of our framework when the retrieval result is perfect.

SMN (Wu et al., 2017) This is a classic work in retrieval-based chatbots, which proposed a sequential matching network to match a response with each utterance on multiple levels of granularity and accumulate the obtained matching vectors with RNN for the final matching score.

MSN (Yuan et al., 2019) This is a recent work in retrieval-based chatbots, which proposed a multi-hop selector network to alleviate the side effect of using unnecessary context utterances. It is one of the most effective methods recently.

3.3 Discrimination Module

In our framework, the discrimination module is used to receive the retrieved high-quality response from the retrieval module, and discriminate the polarity of the response. For the response with satisfied polarity, the module outputs it directly.

Noting that the module handles a classification task, so we can utilize many existing classifiers. In this work, we choose the pre-trained BERT model as our classifier, which has achieved state-of-the-art performances across a variety of NLP tasks.

For the pre-trained BERT model, given a response $R = \{w_1, w_2, w_3, ..., w_n\}$, the input can be expressed as: " [CLS] w_1 w_2 w_3 ... w_n [SEP] ". Following the usual practice, we use the hidden representation for the [CLS] token to represent the response, and then feed it into a softmax layer for classification.

3.4 Rewriting Module

In our framework, the rewriting module is used to receive the response with unsatisfied polarity from the discrimination module, and correct the polarity of response from unsatisfied to satisfied.

Inspired by previous works in style transfer (Li et al., 2018; Sudhakar et al., 2019), we regard the polarity rewriting of the response as a two-stage process: Delete and Generate. The first *Delete* stage employs a pretrained sentiment classification model to delete the affective expressions in the response, and the second *Generate* stage adopts two transformer-based generators to produce a response with satisfied polarity. We introduce the two stages in the following sections.

3.4.1 Delete

In this stage, our goal is to identify and delete the affective expressions in affective responses. For neutral responses, we do nothing at this stage.

Our approach is based on a pretrained sentiment classification model to automatically identify word-level affective expressions. For a sentiment classification model, the affective expressions in a sentence are the key to recognize the polarity of the sentence. Therefore, an intuitive idea is to measure the importance of different words to sentence sentiment classification, and the most important words should be the key affective expressions.

Specifically, we design a word ranking mechanism for identifying word-level affective expressions in the response. We calculate the importance score I_{w_i} for each word w_i in the response R. The method is to remove the word w_i in the response, and compare the target polarity prediction score before and after the deletion, which are $S_e(R_{[w_i]})$ and $S_e(R_{[w/o\ w_i]})$ respectively. The importance score I_{w_i} for each word w_i can be formally defined as follows:

$$I_{w_i} = S_e(R_{[w_i]}) - S_e(R_{[w/o\ w_i]}) \quad (1)$$

We calculate the importance score for each word and choose the top $\lambda\%$ of words as affective expressions. Then, we delete these affective expressions and send the modified response to the next stage.

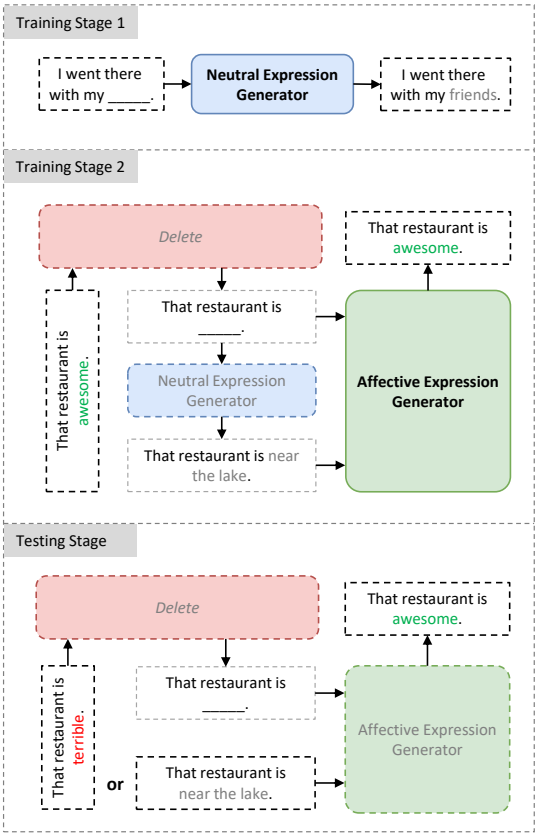


Figure 3: Two training stages and the testing stage of the *Generate* stage in the rewriting module. The blocks with bold font participate in parameter updating.

3.4.2 Generate

In this stage, our goal is to generate a response with a specific polarity. Different from existing works in style transfer (Li et al., 2018; Sudhakar et al., 2019), our polarity rewriting will involve a large number of cases from neutral expression to affective expression, not just the transfer between different affective expressions. An obvious problem is affective responses have affective expressions that can be deleted, but neutral responses have no affective expressions to delete. After the *Delete* stage, although both will become neutral, the sentence distribution of the two is obviously different. If only affective responses can participate in generation training, it will lead to poor performance of the generator for neutral responses.

To address this problem, we propose two generators: neutral expression generator and affective expression generator. We introduce the two generators in the following sections.

Neutral Expression Generator The neutral expression generator is used to complete an incomplete neutral response to a complete neutral response. In the training phase, this generator completes the incomplete neutral responses from the *Delete* stage, which can provide additional training data for the affective expression generator. Thus, the affective expression generator will receive two types of neutral responses at the same time in generation training, which solves the above problem of inconsistent distribution. The architecture of the generator is consistent with Generative Pre-trained Transformer (GPT) (Radford et al., 2018).

Affective Expression Generator The affective expression generator is used to generate a response with satisfied polarity from an incomplete or complete neutral response. In the training phase, the incomplete neutral response is obtained after the *Delete* stage, and the complete neutral response is provided by the neutral expression generator. The architecture of the generator is also consistent with GPT, and we add different special symbols to input for different target polarities to distinguish.

Training and Testing To train the two generators, an affective corpus is required, which contains positive, negative and neutral sentences. The training process consists of two stages, which is shown in Figure 3. In training stage 1, we use neutral sentences to train the neutral expression generator. The input is a processed sentence with $\lambda\%$ words deleted randomly, and the target is the original neutral sentence. In training stage 2, we use affective sentences to train the affective expression generator. The input is an affective sentence, which is processed into an incomplete neutral response and a complete neutral response, and the target is the original affective sentence. In the testing stage, the input is an affective or neutral sentence, and the target is a sentence with specified polarity.

4 Sentimental Douban Conversation Corpus

In this paper, to solve the problem of no off-the-shelf dataset, we annotate the Douban Conversation Corpus (Wu et al., 2017) in terms of sentiment polarity to support the research of obtaining affective response in retrieval-based chatbots.

4.1 Douban Conversation Corpus

This dataset contains open domain multi-turn conversations in Chinese, and it is constructed from Douban group which is a popular social networking service in China. For each dialogue in training and validation sets, the last turn is taken as a positive re-

Table 1: Data statistics for the Douban Conversation Corpus.

	Train	Val	Test
# context-response pairs	1M	50k	10k
# candidates per context	2	2	10
Avg. # turns per context	6.69	6.75	6.45
Avg. # words per utterance	18.56	18.50	20.74

Table 2: Data statistics for the sentiment annotation in the Sentimental Douban Conversation Corpus.

	Manual	Automatic
# contexts	1,400	498,600
# positive utterances	1,981	856,486
# neutral utterances	6,627	2,669,501
# negative utterances	2,104	821,213
# all utterances	10,712	4,347,200

sponse, and another randomly sampled response is taken as a negative response. For each dialogue in test set, it has 10 candidate responses which is collected by an index system and annotated manually. The data statistics are shown in Table 1.

4.2 Sentiment Annotation

As mentioned previously, there is no off-the-shelf dataset to support this task. Such a dataset can not only be used in our framework, but also necessary for existing methods which employs the reranking mechanism. To address this problem, we annotate the Douban Conversation Corpus in terms of sentiment polarity, and obtain a new Sentimental Douban Conversation Corpus.

Specifically, we give annotation guidelines and examples to three human annotators, who then manually annotate sentiment labels for 1,400 dialogues with a total of 10,712 utterances. We extract 1,000 utterances as a shared annotation part of all annotators, and divide the remaining utterances into 3 parts as independent annotation part of each annotator. We measure pairwise inter-annotator agreement among the three annotators in the shared part using Cohen's kappa, and their scores are 0.81, 0.79 and 0.80. For the remaining 498,600 dialogues, we train a classifier using manual annotation data to annotate them automatically. In this work, the manual annotation data is used to train baseline models and our discrimination module, and the automatic annotation data is used to train our rewriting module. Our classifier is a fine-tuned RoBERTa-large model whose pre-training parameters are derived from Chinese RoBERTa (Cui et al., 2020), and obtains the accuracy of 82.79% and the macro-F1 of 79.18% on the divided test set of manual annotation data. A summary of statistics for sentiment annotation is shown in Table 2.

5 Experiments

5.1 Baselines

As mentioned previously, the related research in retrieval-based chatbots is still in the early stage, thus there are very few closely-related baselines. In this paper, we choose two suitable baselines:

Base (w/o. control) This is a basic baseline, which directly outputs the best response matched by the retrieval model without considering target polarity. This baseline represents the ability of the standard retrieval model to obtain affective responses. Note that this baseline only selects responses based on relevance, so it is a very strong baseline in terms of response content quality.

Reranking (Lubis et al., 2019) This baseline is a reranking strategy, which first uses the retrieval model to perform semantic matching on response candidates, and then reranks them according to whether a response satisfies the target polarity. In our experiments, we use the same classifier as our discrimination module. If there are responses satisfying the target polarity, we output the one with the highest semantic matching score. Otherwise, we directly output the best one without considering the target polarity.

5.2 Evaluation Metrics

In this section, we introduce the metrics to evaluate the performance of our proposed framework.

Inspired by related works in generation-based chatbots (Zhou et al., 2018; Song et al., 2019; Shen and Feng, 2020), we perform human evaluation to analyze the quality of the responses from content (Con.), fluency (Flu.) and polarity accuracy (Acc.). First, we randomly sample 100 dialogues from the test set. For each dialogue, we require both positive and negative responses. We present the triples of (context, response, polarity) to three human annotators without order, and they evaluate responses on content, fluency and polarity accuracy independently. Content is measured by a 5-scale rating, which is determined by whether a response is coherent and meaningful for the context. Fluency is measured by a 5-scale rating, which is determined by whether a response is fluent and grammatical.

Table 3: Experimental results on the Sentimental Douban Conversation Corpus. The results are divided into different groups according to different retrieval-based methods, and the comparison of different models is performed within the group. "Con.", "Flu." and "Acc." denote content, fluency and polarity accuracy, respectively.

Group	Model	Positive			Negative			Overall		
		Con.	Flu.	Acc.	Con.	Flu.	Acc.	Con.	Flu.	Acc.
GTM	Base (w/o. control)	5.000	5.000	0.437	5.000	5.000	0.207	5.000	5.000	0.322
	Reranking	3.607	4.730	0.707	3.070	4.663	0.567	3.338	4.697	0.637
	Ours	4.467	4.697	0.803	3.993	4.297	0.723	4.230	4.497	0.763
SMN	Base (w/o. control)	3.787	4.670	0.347	3.743	4.670	0.267	3.765	4.670	0.307
	Reranking	3.430	4.687	0.687	2.973	4.597	0.563	3.202	4.642	0.625
	Ours	3.627	4.457	0.737	3.450	4.227	0.610	3.538	4.342	0.673
MSN	Base (w/o. control)	4.047	4.727	0.337	4.017	4.707	0.257	4.032	4.717	0.297
	Reranking	3.460	4.687	0.687	2.943	4.580	0.567	3.202	4.633	0.627
	Ours	3.830	4.470	0.767	3.623	4.237	0.650	3.727	4.353	0.708

Polarity accuracy is measured by a 2-scale rating, which is determined by whether a response satisfies the target polarity. Note that we do not use automatic metrics because they are not applicable to this task, the detailed explanation can be found in Appendix A.

5.3 Experimental Settings

The architecture and training process of our framework have been introduced in the previous sections. Further training details and hyperparameter values can be found in Appendix B. Our dataset and the implementation for our model are released at https://github.com/luxinxyz/RDR/.

5.4 Overall Results

We compare our proposed framework with the baseline methods, and the experimental results are shown in Table 3. We divide the results based on different retrieval-based models into different groups, and the comparison of the experimental results in each group is fair.

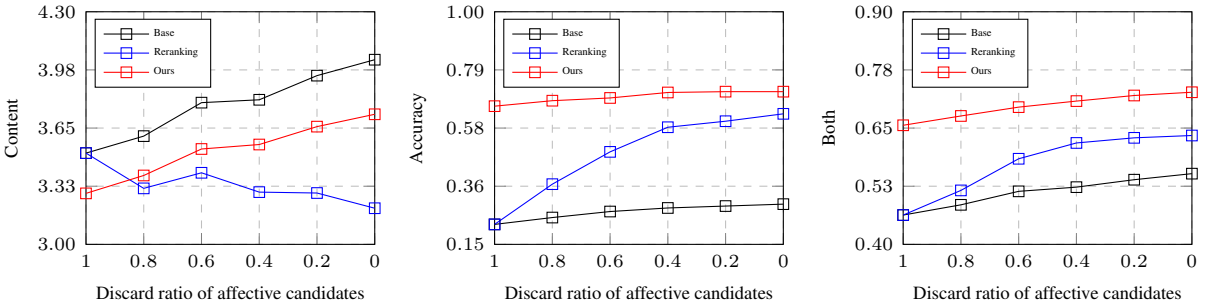
From the perspective of content, *Base (w/o. control)* is the baseline which only considers content without considering polarity, so its content score is the highest among the three methods. Our framework is second only to *Base (w/o. control)* and is significantly better than *Reranking*, which preliminarily illustrates the advantages of our framework in content. From the perspective of fluency, our framework is slightly weaker than *Base (w/o. control)* and *Reranking* because of the modification of the response, but it is also close to the full score. From the perspective of polarity accuracy, our framework is the best among the three methods, which shows the advantages of our framework in polarity accuracy.

Based on the above results, we can see that our framework can obtain affective response better than the baseline methods, especially on the basis of ensuring polarity accuracy, effectively avoiding the low-quality response problem of the reranking mechanism. In addition, the reproduced results of the retrieval-based methods can be found in Appendix C, and more detailed response examples can be found in Appendix E.

5.5 Analysis

5.5.1 Impact of Affective Candidate Size

We analyze the impact of affective candidate size to further explain the problem of the Retrieve-and-Rerank framework and the advantage of our Retrieve-Discriminate-Rewrite framework. Specifically, we control the ratio of affective responses in the response repository by discarding them to simulate retrieval-based dialogue systems with different level affective information, and then plot the performance trends of different methods on content, polarity accuracy and the mean of both after normalization. All experiments are under the MSN settings, and the results are shown in Figure 4. Under normal circumstances, with the increase of affective candidates(the decrease of discard ratio) in dialogue systems, the content score should gradually increase, just like *Base (w/o. control)* and our framework. However, the content score of *Reranking* gradually decreases, which confirms low-quality response problem of the Retrieve-and-Rerank framework we mentioned in the introduction. From the perspective of polarity accuracy, our



(a) Content Performance (b) Affect Performance (c) Mean Performance

Figure 4: Analyses of the impact of affective candidate size in terms of content and affect on the Sentimental Douban Conversation Corpus.

Table 4: Analyses of the impact of discrimination module on the Sentimental Douban Conversation Corpus.

Models	F1	Con.	Flu.	Acc.
Ours-Discrim-CNN	0.601	4.197	4.490	0.733
Ours-Discrim-BiLSTM	0.636	4.193	4.478	0.718
Ours-Discrim-BERT	0.791	4.230	4.497	0.763

framework can always maintain a high level. Finally, considering the overall results of content and polarity accuracy, our framework is better than the other two methods, which proves the effectiveness of our framework.

5.5.2 Impact of Discrimination Module

We analyze the impact of our discrimination module on the final performance. Specifically, we replace the classifier of our discrimination module from BERT to different architectures, such as CNN and BiLSTM, and then explore the relationship between the performance of our discrimination module and the final performance. All experiments are under the GTM settings, and we use macro-F1 to evaluate the classifier performance. As presented in Table 4, the best performance classifier corresponds to the best final performance, which illustrates the importance of a good discrimination module in our framework.

5.5.3 Analysis of Rewriting Module

We analyze the rewriting module in our framework. Specifically, we reproduct a style transfer model named *DeleteRetri* (Li et al., 2018) to compare with our proposed rewriting module. We choose this model because it also includes the process of deletion and generation, but has no special design for neutral response. And in order to verify the ability to process neutral response, we evaluate the polarity accuracy when the input of these models is affective (Acc-A.) and neutral (Acc-N.) respectively. All experiments are under the GTM settings, and the results are shown in Table 5. From the table, we observe that the content and fluency of the two models are similar, but the polarity accuracy of our rewriter is significantly better. *DeleteRetri* has the problem that the performance of neutral input is significantly lower than that of affective input while our rewriter does not have such a problem, which shows the effectiveness of our improvement. We also compare with other style transfer models, and the results can be found in Appendix D.

Table 5: Analyses of the impact of rewriting module on the Sentimental Douban Conversation Corpus.

	Models	Con.	Flu.	Acc.	Acc-A.	Acc-N.
Positive	DeleteRetri	4.467	4.553	0.707	0.783	0.635
	Ours	4.467	4.697	0.803	0.809	0.799
Negative	DeleteRetri	4.147	4.433	0.507	0.525	0.491
	Ours	3.993	4.297	0.723	0.716	0.730
Overall	DeleteRetri	4.307	4.493	0.607	0.656	0.563
	Ours	4.230	4.497	0.763	0.762	0.764

6 Conclusion

In this paper, we propose a Retrieve-Discriminate-Rewrite framework for obtaining affective response in retrieval-based chatbots, which solves the problem of low-quality responses in the Retrieve-and-Rerank framework. Our framework contains three components: retrieval module, discrimination module and rewriting module, which can preferentially select high-quality candidate response and rewrite the response whose affect is discriminated to be unsatisfied. Considering the lack of necessary dataset in this field, we further annotate and publish a Sentimental Douban Conversation Corpus. The empir-

ical studies show that our framework outperforms competitive baselines, and extensive analyses further proves the effectiveness of our framework.

Acknowledgements

This work was supported by the National Key R&D Program of China via grant 2018YFB1005103 and National Natural Science Foundation of China (NSFC) via grant 61632011 and 61772153. We thank the anonymous reviewers for their insightful comments and suggestions. We also thank Weixiang Zhao, Wenjia Yi, Song Chen, Pai Peng, Hao Yang, Xin Zhang and Yanyue Lu for their help in this work.

Ethical Considerations

In this section, we address relevant ethical considerations that were not explicitly discussed in the main body of our paper.

Intended Use The reported technique is intended for building affective chatbots used in daily life. We anticipate that this technique will significantly promote affective communication and enhance user satisfaction during human-computer interactions, which is an enhancement to existing chatbots.

Potential Misuse In some cases, our proposed model may produce effects similar to mental health support. This may mislead users that the model has professional psychotherapeutic capabilities, leading to misuse. In fact, this model is not developed from the perspective of professional psychology applications. Applying it to professional-level mental health support is extremely risky, and in extreme cases it may cause harm to users. We reiterate once again that the reported technique is only intended for building affective chatbots used in daily life.

Failure Modes The main failure mode is that the model may learn some bad expressions in the training data which are harmful to users. Based on the consideration of compatibility with existing works, we performed additional annotations on a widely used dataset and trained the model based on it. This dataset is an early classic dataset which does not represent current norms and practices, so there is indeed a possibility of harmful responses(but actually very few), which may involve offensive speech, hate speech, etc. In order to reduce this risk, one idea is to clean the harmful responses in the dataset, and the other is to detect the harmfulness of the results output by the model. Both of these can be achieved based on some recent works on offensive speech detection (Ranasinghe and Zampieri, 2020) or hate speech detection (Vidgen et al., 2021).

In addition, in order to provide an intuitive reference for users of the model, we conducted an empirical evaluation of the harmfulness of the model. Specifically, we randomly sampled 1,000 responses output by the model and asked three human annotators to evaluate whether the responses might make users uncomfortable. The evaluation results show that only 19 responses will make users slightly uncomfortable, and there are no responses that make users seriously uncomfortable. This result is not enough to completely eliminate concerns, but in a sense, it shows the actual performance of the model trained on the dataset.

References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of ACL 2019*.

Kristie S Fleckenstein. 1991. Defining affect in relation to cognition: A response to susan mcleod. *Journal of Advanced Composition*, pages 447–453.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of AAAI 2018*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *Proceedings of ICLR 2019*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura. 2019. Positive emotion elicitation in chat-based dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):866–877.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

Timo Partala and Veikko Surakka. 2004. The effects of affective interventions in human-computer interaction. *Interact. Comput.*, 16(2):295–309.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *Proceedings of ACL 2018*.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users' affective states. *Appl. Artif. Intell.*, 19(3-4):267–285.

Lisong Qiu, Yingwai Shiu, Pingping Lin, Ruihua Song, Yue Liu, Dongyan Zhao, and Rui Yan. 2020. *What If Bots Feel Moods?*, page 1161–1170. Association for Computing Machinery, New York, NY, USA.

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

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. 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 1577–1586, Beijing, China. Association for Computational Linguistics.

Lei Shen and Yang Feng. 2020. CDL: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 556–566, Online. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of NIPS 2017*.

Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4382–4388. International Joint Conferences on Artificial Intelligence Organization.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of ACL 2018*.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3801–3807. International Joint Conferences on Artificial Intelligence Organization. Main track.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 111–120, Hong Kong, China. Association for Computational Linguistics.

JianHua Yuan, Yang Wu, Xin Lu, YanYan Zhao, Bing Qin, and Ting Liu. 2020. Recent advances in deep learning based sentiment analysis. *Science China Technological Sciences*, pages 1–24.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *CoRR*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

A Additional Description of Evaluation Metrics

Note that we do not use automatic metrics to evaluate affective responses, because existing automatic metrics are not suitable. For the content perspective, the automatic metrics such as MAP, MRR, P@1 and R_n@k for retrieval evaluation are not suitable for the responses that are not in the response repository, and the automatic metrics such as Perplexity, BLEU scores and Embedding scores for generation evaluation are not suitable for the responses retrieved from the response repository. Therefore, although we can use automatic metrics to test our framework, we cannot form a valid comparison with the baseline models. For the polarity accuracy perspective, we can not use a sentiment classifier to evaluate affective responses, because the baseline methods and our framework already rely on sentiment classifiers, especially $Reranking$ completely relies on the results of the sentiment classifier. For the above reasons, we only perform human evaluation to analyze the quality of affective responses.

B Training Details

For our retrieval module, we use open-source codes[1,2] provided by the authors to reproduct SMN (Wu et al., 2017) and MSN (Yuan et al., 2019) with the same settings as original papers.

For our discrimination module, we use manual annotation data of the dataset for training. We initialize the module with the pre-training parameters provided by Chinese RoBERTa (Cui et al., 2020), and our implementation is based on the PyTorch implementation of BERT-large[3]. We use Adam (Kingma and Ba, 2015) as optimizer with a learning rate of 1e-5 and a batch size of 16, and use the linear learning rate decay schedule with warmup over 0.1. We set the maximum number of epochs to 5 and select the model with the best performance on the validation set. The average runtime is 3 hours on a Tesla V100 32GB GPU machine.

For our rewriting module, in the *Delete* stage, we use manual annotation data of the dataset for training, and the settings are similar to the discrimination module. In the *Generate* stage, we use automatic annotation data of the dataset for training.

[1]https://github.com/MarkWuNLP/MultiTurnResponseSelection

[2]https://github.com/chunyuanY/Dialogue

[3]https://github.com/huggingface/transformers

Table 6: Results of different retrieval-based methods on the Douban Conversation Corpus.

Models	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5
TF-IDF†	0.331	0.359	0.180	0.096	0.172	0.405
CNN†	0.417	0.440	0.226	0.121	0.252	0.647
LSTM†	0.485	0.527	0.320	0.187	0.343	0.720
SMN†	0.529	0.569	0.397	0.233	0.396	0.724
MSN‡	0.587	0.632	0.470	0.295	0.452	0.788
SMN	0.541	0.583	0.391	0.230	0.409	0.779
MSN	0.581	0.629	0.459	0.285	0.453	0.792
GTM	1.000	1.000	1.000	1.000	1.000	1.000

Table 7: Results of different style transfer models on the Sentimental Douban Conversation Corpus.

Models	Con.	Flu.	Acc.
CrossAlign (Shen et al., 2017)	4.070	4.102	0.370
StyleTransformer (Wang et al., 2019)	4.177	4.232	0.485
DualRL (Luo et al., 2019)	4.262	4.337	0.555
StyIns (Yi et al., 2020)	4.250	4.295	0.523
Ours-Rewriting-Module	4.230	4.497	0.763

For our two generators, we use the same structure as Generative Pre-trained Transformer (GPT) (Radford et al., 2018), and our implementation is based on the PyTorch implementation of GPT[4]. For both models, we use Adam (Kingma and Ba, 2015) as optimizer with a learning rate of 1e-4 and a batch size of 256, and use the linear learning rate decay schedule with warmup over 0.1. For the neutral expression generator, we train the model for 60 epochs on 8 Tesla V100 16GB GPU machines, which takes about 40 hours. For the affective expression generator, we set the maximum number of epochs to 10 and select the model with the best performance on the validation set. The average runtime for the generator is 20 hours on a Tesla A100 40GB GPU machine. Based on the model performance on the validation set, the λ is set to 25 in the *Delete* and *Generate* stage. In the testing phase, we using the Nucleus Sampling (Holtzman et al., 2020) with a threshold 0.9 and temperature 0.7 to decode responses.

C Retrieval Model Performance

We reproduct some retrieval-based methods as the retrieval module of our proposed framework, and we use the automatic metrics such as MAP, MRR, P@1 and R_n@k for retrieval evaluation to evaluate these methods.

The results of retrieval-based methods are shown in Table 6. The results marked by † are from the

[4]https://github.com/thu-coai/CDial-GPT

paper of SMN, and the results marked by ‡ are from the paper of MSN. From the table, we observe that our results reach the level of original papers. In the above experiments, we directly use these models as our retrieval module.

D Style Transfer Experiment

In addition to *DeleteRetri* (Li et al., 2018) , we also reproduce other style transfer models (Shen et al., 2017; Wang et al., 2019; Yi et al., 2020; Luo et al., 2019) to compare with our proposed rewriting module. All experiments are under the GTM settings, and the results are shown in Table 7. From the table, we observe that the content and fluency of these models are similar, but the polarity accuracy of our rewriting module is the best among these models. The main reason for the result is our model adopts a rewriting architecture that can achieve fine-grained control of affect, which is more suitable for this task than other models.

E Sample Affective Responses

We show some examples obtained from baselines and our framework under the GTM settings. As presented in Table 8, $Base(w/o.control)$ can output high-quality but polarity unsatisfied response in most cases, and $Reranking$ can output polarity satisfied but poorly relevant response in most cases, while our framework can always output polarity satisfied high-quality response by the discriminate-and-rewrite mechanism. These cases show that our proposed framework is effective.

Table 8: Some sample responses from Base(w/o.control), Reranking and Ours (original Chinese and English translation). Affective expressions are marked with colors.

Dialogue History	A: 泰拳爱好者	A: I am a Muay Thai lover.
	B: 女生对泰拳感兴趣的怎么办？	B: I am a girl, what should I do if I am interested in Muay Thai?
	A: 我练了半年泰拳，可以交流下	A: I have practiced Muay Thai for half a year, we can talk about it.
	B: 太好了，回家加你	B: Great, I will add you when I get home.
Affect Label	消极	*Negative*
Base (w/o. control)	A: 额，我一般晚上才上	A: Um, I usually only go online at night.
Reranking	A: 嗯呢，机器人比较无解的	A: Well, the robot is relatively **unsolvable**.
Ours	A: 额，我一般晚上上班，累死	A: Um, I usually go to work at night and I'm **exhausted**.
Dialogue History	A: 没有恋爱但是失恋了，有没有这个可能性?	A: Is there such a possibility that I am not in love but I am lovelorn?
	B: 各种暗恋再失恋	B: There are many cases of lovelorn after secret love.
	A: 那我悲剧了	A: That's a tragedy for me.
	B: 你会很强大的	B: You're going to be strong.
Affect Label	消极	*Negative*
Base (w/o. control)	A: 会的会的	A: Yes, I will.
Reranking	A: 毕业生找不到事情做，诶，下雨天又不能粗门	A: I can't find a job after graduation. **Alas**, it's raining today and I can't go out.
Ours	A: 会的会的，心痛	A: Yes, I will, but it's **heartbreaking**.
Dialogue History	A: 好准的星座啊	A: Astrology is really accurate.
	B: 我都被自己的星座感动了	B: Yes, I am moved by my astrology-based personality test result.
	A: 你是什么?	A: What star sign are you?
	B: 双子	B: Gemini.
	A: 双子蛮好的，我很喜欢	A: Gemini is good. I like it very much.
Affect Label	消极	*Negative*
Base (w/o. control)	B: 吼吼，可有人不喜欢	B: Ho Ho, but some people **don't like** it.
Reranking	B: 吼吼，可有人不喜欢	B: Ho Ho, but some people **don't like** it.
Ours	B: 吼吼，可有人不喜欢	B: Ho Ho, but some people **don't like** it.
Dialogue History	A: 阳澄湖大闸蟹，是不是就是温州人讲的田丝儿啊?	A: Yangcheng Lake hairy crabs, is it the 'Tian Si' that Wenzhou people talk about?
	B: 我觉得还是蝤蠓最好吃啊	B: I think the 'You Meng' are the best.
	A: 是神马东东啊，我都没听过	A: What is that? I've never heard of it.
	B: 你肯定吃过	B: You must have eaten it.
Affect Label	积极	*Positive*
Base (w/o. control)	A: 也许吧，吃的时候不认识它	A: Maybe. I didn't know it when I ate it.
Reranking	A: 和大龙比怎么样阿，我觉得余总还是很巴适的	A: How about comparing with Dalong? I think Mr. Yu is very **friendly**.
Ours	A: 也许吧，期待吃的时候开心	A: Maybe. I look forward to being **happy** when I eat it.
Dialogue History	A: 离开淄博回到潍坊	A: I left Zibo and returned to Weifang.
	B: 淄博人，马上要去潍坊	B: I'm from Zibo. I'm going to Weifang.
	A: 哈哈，什么时候?	A: Ha ha, when?
	B: 4月初，求接待	B: In early April, could you receive me?
Affect Label	积极	*Positive*
Base (w/o. control)	A: 哈哈，加我提前告诉下行程	A: **Ha ha**, you can add me and tell me your itinerary in advance.
Reranking	A: 哈哈，加我提前告诉下行程	A: **Ha ha**, you can add me and tell me your itinerary in advance.
Ours	A: 哈哈，加我提前告诉下行程	A: **Ha ha**, you can add me and tell me your itinerary in advance.