Predicting Users' Negative Feedbacks in Multi-Turn Human-Computer Dialogues *

Xin Wang[1], **Jianan Wang**[2], **Yuanchao Liu**[1], **Xiaolong Wang**[1], **Zhuoran Wang**[3] and **Baoxun Wang**[3]
[1]Harbin Institute of Technology, Harbin, China
[2]Shanghai Jiao Tong University, Shanghai, China
[3]Tricorn (Beijing) Technology Co., Ltd, Beijing, China
[1]{xwang,lyc,wangxl}@insun.hit.edu.cn
[2]{wangjianan}@sjtu.edu.cn
[3]{wangzhuoran, wangbaoxun}@trio.ai

Abstract

User experience is essential for human-computer dialogue systems. However, it is impractical to ask users to provide explicit feedbacks when the agents' responses displease them. Therefore, in this paper, we explore to predict users' imminent dissatisfactions caused by intelligent agents by analysing the existing utterances in the dialogue sessions. To our knowledge, this is the first work focusing on this task. Several possible factors that trigger negative emotions are modelled. A relation sequence model (RSM) is proposed to encode the sequence of appropriateness of current response with respect to the earlier utterances. The experimental results show that the proposed structure is effective in modelling emotional risk (possibility of negative feedback) than existing conversation modelling approaches. Besides, strategies of obtaining distance supervision data for pre-training are also discussed in this work. Balanced sampling with respect to the last response in the distance supervision data are shown to be reliable for data augmentation.

1 Introduction

As an ideal interaction mode, the human-computer conversation technology has gradually been conducted into the practical systems, among which the automatic agents adopting the task-oriented conversation and open-domain chatting abilities have developed rapidly and even come to commercial application stage (see Duer[1] and XiaoIce[2]).

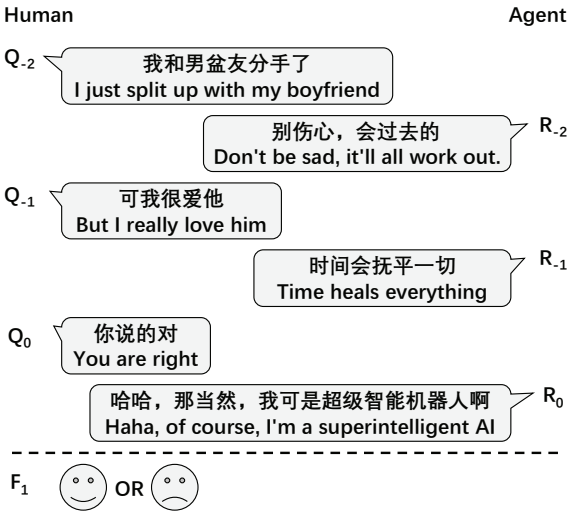

Figure 1: A example of multi-turn human-computer dialogue. The responses are readable and relevant to the corresponding queries, but the user in bad mood probably feel antipathy towards the computer's cheer.

Basically, the essence of such commercial dialogue agents is keeping user active, and for this purpose, it is critical to improve users' satisfaction in non-task oriented chatting services.

The reasons of users' displeasure lies in the sessions. As shown in Figure 1, the emotional conflict between the last agent's response and the user's mood leads to the dissatisfaction. Therefore, the negative feedback from user could be predicted according to the context utterances. Indeed, such prediction is of great necessity for improving users' satisfaction, since it is possible to address problems within systems only if there exist methodologies to locate problems and summarize reasons behind. Ideally, if a dialogue agent anticipates the occurrence of the user's negative feedback, it is able to avoid such situations by taking any possible actions, e.g., switching to an-

*The work was done when the first author was an intern at Tricorn (Beijing) Technology Co., Ltd.

[1]http://duer.baidu.com/

[2]http://www.msxiaoice.com/

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 713–722,
Taipei, Taiwan, November 27 – December 1, 2017

other topic actively. Moreover, user satisfaction can be taken as a metric for evaluating the quality of a given dialogue session and the overall performance of a dialogue agent. In practice, user satisfaction could also be considered as the additional criterion for response selection or generation, which tend to intuitively take semantic relevance oriented features for model training. The adoption of user satisfaction is possible to provide a different view to optimize the models, so as to further improve user experience along the road beyond relevance, e.g., avoiding the responses that are relevant but lack of sociality (Higashinaka et al., 2015).

It is not a trivial task to predict negative feedbacks in the conversation flows between human being and dialogue agents. Generally, people tend to not express their dissatisfaction explicitly, thus, there are generally no clear signals before users turn angry and terminate dialogues, or people continue the conversations although they are not satisfied already. Apparently, it is unwise to introduce rating or other explicit feedback mechanisms into the dialogue flows considering the user experience issues. Meanwhile, this problem cannot be covered by classical sentiment analysis task because users' sentiment intention tends to be not obvious, and more importantly, the facts causing negative feedbacks are much more complicated than sentiment polarities as shown by Figure 1, in which the appropriateness of a certain response might be decided not only by itself, but also by the context.

In order to estimate the risk of dissatisfaction occurring in the human-computer dialogue sessions, in this work, we explore the feasibility of predicting users' emotional negative feedback caused by the dialogue agents' replies based on the dialogue contexts. To our knowledge, this is the first work attempting to discover the implicit factors causing users' dissatisfaction in dialogue agents' logs with deep learning models. Noticing that the occurrence of negative feedbacks depends on a complicated semantic mechanism and conversational contexts play an important role in this issue, this paper proposes to address the problem by learning to represent the possible determinant with different models. Especially, the proposed architecture based on Gated Convolutional Recurrent Neural Networks (GCRNN) is used to represent sequence of relations between the last response and the earlier utterances. Experimentally, it outperforms existing conversational models, which indicates that the sequence of relation between utterances encodes the possibility of user's dissatisfaction. Besides, data augmentation with distance supervision method is also discussed in this work.

2 Related Work

2.1 Emotion prediction

Predicting sentiment category of text has been extensively studied. Most works focus on the sentiment orientations expressed by the writers in movie/product reviews or tweets (Pang et al., 2002; Hu and Liu, 2004; Go et al., 2009). However, the reader's emotion is not always consistent with that of the writer's (Yang et al., 2007). Thus, Lin et al. (2007) explore to predict the feelings that readers may have after reading particular articles. However, in this dissatisfaction prediction task, the user is not only the reader of the session text, but also the writer of some utterances. Therefore, some clues of the particular user's emotion may be contained in the context.

Modelling emotion in human-human conversation has been explored (Herzig et al., 2016; Tokuhisa and Terashima, 2006). However, triggers for negative emotion in human-computer dialogue might be different (e.g., low readability or relevance). The works of Tokuhisa et al. (2009) and Yu et al. (2016)'s analysed the emotion of a particular utterance in human-computer dialogue based on the textual features containing in the very sentence. Different from their studies analysing the explicit textual feedback, this work predicts the impending emotion based on the context because the cause of emotion of readers contains in the existing text (Li and Xu, 2014).

2.2 Conversation Modelling

Traditional conversation modelling mainly focuses on the one-turn conversations (aka. message-response pairs) (Banchs and Li, 2012; Ameixa et al., 2014; Ritter et al., 2011; Ji et al., 2014), while recent works show more interest in multi-turn dialogues.

The generation-based approaches model the context and generate the responses at the same time (Vinyals and Le, 2015), while the retrieval based studies model the sessions after knowing all utterances, which is more relevant to this work. Xu et al. (2016) represent sequence of utterances with recurrent neural networks (RNNs). The topic

or intention in a dialogue session is relatively constant. In this perspective, all the utterances in the same session is homogenous and could be composed within RNNs. However, the influences of user's queries and the agent's responses are different in predicting user's emotion. Therefore, a targeted structure considering such heterogenicity is proposed in this work.

Wu et al. (2016) represent the relevance between utterances with CRNN architecture. Different from their work focusing on word-level matching with attention pooling on the convolutional result, we leverage a gate operation to simulate the sentence-level interaction.

3 Predicting Methodology

The task of predicting impending dissatisfaction could be formulated as: given the existing utterances (EU) that contain no agent-cause dissatisfaction, predicting the agent-cause dissatisfaction $D_1 \in \{0, 1\}$ of user at impending turn ($r = 1$), given the existing utterances (EU) that contain no agent-cause dissatisfaction.

$$EU = \{Q_{-n+1}, R_{-n+1}, ..., Q_{-1}, R_{-1}, Q_0, R_0\} \tag{1}$$

where Q_{-n} and R_{-n} respectively represent the user's query and computer response n round before current turn. For this work focusing on the agent-caused dissatisfaction, those queries Q_{-n} with negative emotion not related to the robot are not considered as negative feedback.

There are several possible factors influencing the emotion of the users, such as (1) the last response of the robot R_0, (2) the relation of Q_0 and R_0, (3) the sequence of context in the conversational sessions EU and (4) the sequence of relations between R_0 and the other utterances $UE - \{R_0\}$. In this session, we will discuss the factors above and learn the representation of them with deep neural network.

3.1 Utterance Modelling

The last response of the robot R_0 is the most straightforward factor that may cause the antipathy towards the agent. Predicting the negative emotions according to the latest response can be considered as reader-side emotion classification. In this work, such single sentence is modelled with convolutional neural network (CNN) with max pooling, and then classified in the full-connected softmax layer (Kim, 2014). The illustration of the utterance model for R_0 is shown in Figure 2(b).

In this paper, all representations of utterances are attracted with CNN structure described in Figure 2(a). An n-word utterance can be represented as:

$$\mathbf{e}_{1:n} = \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus ... \oplus \mathbf{e}_n \tag{2}$$

where $\mathbf{e}_n$ is the embedding of nth word in the utterance and $\oplus$ refers to concatenation operator. In the convolutional process, the word window starting with the ith word and scanned by a s-width filter j can be represented $\mathbf{e}_{i:i+s-1}$. And the activations corresponding with filter j in convolutional layer can be computed as:

$$c_i^j = f\left(\mathbf{w}^j \cdot \mathbf{e}_{i:i+s-1} + b^j\right) \tag{3}$$

where f is the non-linear activation function (ReLU is utilized in this work). And $\mathbf{w}^j$ and b^j represent the weight matrix and bias respectively. Finally, max-over-time pooling is leveraged. The activations corresponding with filter j in the pooling layer can be computed as:

$$\hat{c}^j = max\left\{c_1^j, c_1^j, ..., c_{n-s+1}^j\right\} \tag{4}$$

3.2 Utterance Pair Modelling

Recent works improve response ranking by model the semantic matching of query-response pairs (Qiu and Huang, 2015; Yin et al., 2015). The assumption implied in these works is that user experience is influenced by the relation of Q_0 and R_0. We model such relation with Architecture-I in Hu et al.'s (2014) work, where the representations of the query and the response are learned with two CNNs respectively and the concatenation of the representations is used as input of a multi-layer perceptron (MLP) classifier that measures the appropriateness.

3.3 Utterance Sequence Modelling

As shown in example in Figure 1, the latest response is active and related to the query, but may not appropriate in the context. Recent works encode the sequence of utterances with recurrent neural network based encoder-decoder to generate responses (Serban et al., 2016; Shang et al., 2015). However, in this work, the prediction is made with all existing utterances being known. The convolutional recurrent neural network has been proven to be effective in encoding the sequence of representations of text (Kalchbrenner and Blunsom, 2013;

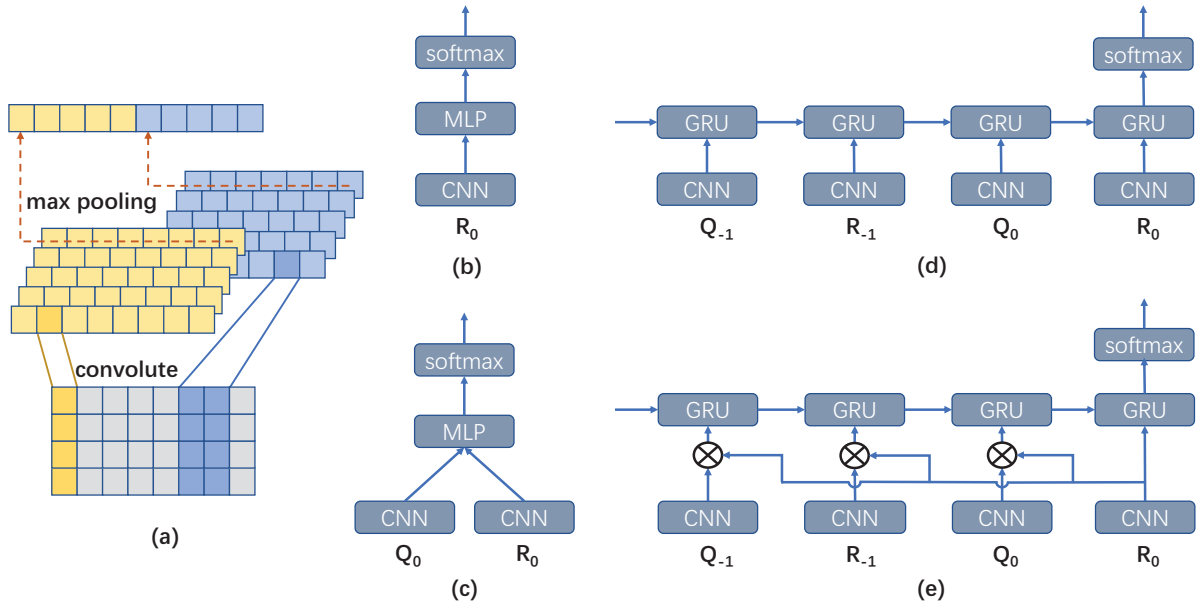

Figure 2: Structures for modelling different influential factors of negative feedbacks.

Li et al., 2016; Zhou et al., 2015). Thus, in this work, the sequence of existing utterances EU is modelled with CRNN, and the output of last time step R_0 is considered as the final representation of the sequence to be input to softmax classifier. The structure is shown in Figure 2(d), and Gated Recurrent Units (GRUs) are used in the structure.

A GRU stores context information in the internal memory structure. It performs comparably with long short-term memory (LSTM) and has lower complexity (Chung et al., 2014). There are two gates in the jth GRU structure, the update gate z_t^j and reset gate r_t^j, both gates are decided by the current input $\mathbf{x}_t$ and previous hidden activation $\mathbf{h}_{t-1}$:

$$z_t^j = \sigma\left(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1}\right)^j \quad (5)$$

$$r_t^j = \sigma\left(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1}\right)^j \quad (6)$$

where W and U is the weight matrices, while σ refers to the sigmoid function. The hidden activation h_t^j of the GRU at time t can be computed as:

$$h_t^j = \left(1 - z_t^j\right) h_{t-1}^j + z_t^j \tilde{h}_t^j \quad (7)$$

where h_{t-1}^j refers to the hidden activation of previous time step and $\tilde{h}_t^j$ is the current candidate activation:

$$\tilde{h}_t^j = tanh\left(W \mathbf{x}_t + U\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right)\right)^j \quad (8)$$

GRUs compose the current and previous information with the gated units, and store the sequence representation in the memory.

3.4 Relation Sequence Modelling

Different from the intentions or topics of utterances being relatively constant in dialogue sessions, emotional influence of each utterance varies. For instance, the queries and responses are heterogeneous in a session. Queries are expressions of dissatisfaction, while responses are the reason of displeasure. The relations between particular response and context queries (or other responses) influence the user's emotion.

As shown in the example in Figure 1, the last response R_0 conveys conflicting emotion with the earlier utterances Q_{-2}, R_{-2} and Q_{-1}. The human user is probably unprepared for such rapid change in emotion and exhibits dissatisfaction. The consistency of mood is a kind of relation between sentences and the sequence of such consistency between sentence pairs can be treated as the emotional consistency of the whole conversation.

However, traditional RNNs are not adequate to represent such consistency. Therefore, we attempt to model the sequence of relation between utterances. As mentioned in the beginning of Section 3, the existing utterances EU contain no agent-cause dissatisfaction, which means Q_0 and the utterances before Q_0 are not the direct causes negative feedback. Thus, we only focus on the relation between the last response and the earlier utterances. The architecture is shown in Figure 2(e), the representation of earlier utterances $\mathbf{x}_t$ $(t \neq 0)$

are gated by representation of R_0:

$$\mathbf{x}_t = \mathbf{c}_t \odot \mathbf{m}_t \tag{9}$$

where $\mathbf{c}_t$ is the output of convolutional neural network sentence model. And the matching gate $\mathbf{m}_t$ is influenced by the particular utterance $\mathbf{c}_t$ and the latest response $\mathbf{c}_0$.

$$\mathbf{m}_t = \sigma\left(W_m \mathbf{c}_t + U_m \mathbf{c}_0\right) \tag{10}$$

While the input of the last time step is the representation of R_0 itself ($\mathbf{x}_0 = \mathbf{c}_0$). Gating operation has been shown effective in further mapping abstract feature of convolutional result by involving additional information (Wang et al., 2015; Dauphin et al., 2016). With such structure, the emotional consistency of utterances could be extracted and the influence of latest response on negative feedback could be encoded.

4 Experiments

4.1 Dataset

The anonymized multi-turn dialogue session data is provided by a Chinese commercial intelligent agent service. There are 2 million sessions in the dataset, most of which contain task-oriented dialogues. However, we focus on those only including chat, and the amount of such pure chat sessions is 260,867.

As described in the introduction, the task is to predict the impending dissatisfaction given $n + 1$ round context. Thus, a sample in the dataset should contains utterances and label of following emotional polarity. In fact, the lengths of human-computer dialogue sessions vary within relatively wide range. To eliminate the influences from session length, n is set to 2 in this work. The appearance of dissatisfaction will be predicted based on 3 turns (0, 1 and 2) of dialogue (as shown in Figure 1). 40,000 of the non-task-oriented sessions are manual annotation to construct the data set. If a session has a negative feedback, the 3 turns of utterances before this feedback will be treated as positive (with dissatisfaction) sample. Otherwise, if there's no negative feedback, we randomly select continuous 3-turn utterances as a negative (without dissatisfaction) sample.

Two experienced annotators (long term employed for text annotation) are scheduled to label the sessions independently. If disagreement appears, a third senior annotator is invited to decide the final tag. Finally, 30,034 sessions meeting the requirements (non-negative sessions or negative sessions with 3 three turns or more utterances before the negative expressions) are used as gold-standard dataset. Some statistical information of the dataset is shown in Table 1.

Category	Amount
total sessions	2,000,000
pure char sessions	260,867
sessions for manually labelling	40,000
two annotators agreement	28,651
the third annotator decision	11,349
gold standard	30,039
negative in gold standard	17,618
positive in gold standard	12,421

Table 1: Statistical information of the dataset.

4.2 Pre-training

4.2.1 Fragment Extraction

The manually labelled gold standard dataset might be insufficient for learning deep neural models. Thus, a distance supervision strategy is designed to obtain augmented data. We summarize 56 patterns of highly probable negative expressions as strict patterns (SP) and 86 patterns of possible or ambiguous negative feedbacks as uncertain patterns (UP). It is noted that the SP is a subset of UP. The sessions containing no utterances matching UP are considered as non-negative. While the utterances (1)containing any SP and (2)with no utterances in the above 3 turns match any UP are treated as negative feedbacks and the fragments are tagged as dissatisfaction ones. In this way, both positive and negative samples of dissatisfaction are automatically detected.

Besides the pure chat sessions, those task-oriented ones also contain multi-turn chat fragment. Therefore, the augmentation strategy is carried out on all available sessions and obtains 1,612,426 distance supervised labelled fragments. It is worthy to note that there may be multiple available fragments in a single session, those fragments (without overlap) are all extracted in the augmentation process.

4.2.2 Balanced Dataset Construction

In fact, the extraction strategy above may lead to a different distribution with the real-world data. Taking the cheerful response R_0 in Figure 1 as an example, most users in bad mood would be upset after the agents reply with such cheer. These users

tend to express dissatisfaction towards the agents and the dissatisfactions are detected with designed patterns. While in the situations where the users are happy, the cheerful response result in a virtuous cycle and may not be selected as R_0 (may be selected as R_{-1} or R_{-2}). Therefore, such cheerful response is likely to closely related to the dissatisfaction.

To avoid such false association rules, we select the same number of positive and negative samples if R_0 is the same and obtain 335,314 fragments as balanced distance supervised labelled data.

4.3 Experimental Settings

All the neural network models are implemented with TensorFlow toolkit[3]. The max length of the input sentence is set to 10 and all sentences are padded to the max length with zero vectors. 32 filters are used for each filter size, while the sizes of word embeddings, hidden layer in RNN and full-connected layer are all set to 64.

The weights between full-connected layers are initialized with Xavier initializer (Glorot and Bengio, 2010), while the weights and biases in the convolutional layer are initialized with random numbers on uniform distribution $\mathcal{U}\{-0.2, 0.2\}$. Word embeddings are randomly initialized with uniform distribution $\mathcal{U}\{-0.1, 0.1\}$ and fine-tuned during training.

Batch learning is conducted with a batch size of 500. The learning rate of the training process is 0.001 while that of pre-training process is 0.005. 10-fold cross validation are implemented with 80% data as training set, and validation and test set divide equally the rest 20% samples. Early stopping is carried out on validation set during training. Training process stops when there's no better validation result within 5 epochs.

4.4 Competitor Models

SVM-R_0: Support vector machine (SVM) are wildly used as classifiers for sentiment analysis tasks (Pang et al., 2002). In this work, TF-IDF features based on uni-grams in R_0 are involved to build baseline model.

SVM-Q_0R_0 and **SVM-EU:** To make use of more context information, we involve Q_0-R_0 pair by connecting them into a whole and uni-gram TF-IDF features of the connection result are used as input of a SVM classifier. In the same way, the all sentences in EU are also used in the SVM model.

UM-Q_0R_0 and **UM-EU:** Similar to SVM, utterance model (UM) shown in Figure 2(b) is also designed to analyse a single sentence (or document). Thus, we leverage connection results of Q_0-R_0 pair and EU to introduce context utterance.

UPM-EU: Besides Q_0-R_0 pair, the utterance pair model (UPM) could also be used to encode all sentences in EU. Each sentence is modelled by CNN respectively and the representations are concatenated in the hidden layer.

4.5 Experimental Results

4.5.1 Comparison with Baselines

We compare the models corresponding to the factor assumptions describe in section 3, including utterance model (UM), utterance pair model (UPM), utterance sequence model (USM) and relation sequence model (RSM) with the competitor systems. The numbers in Table 2 show the proportion of the particular model accurately predicting the emotional polarities. The neural models are pre-trained with balanced distance supervised labelled data, and tuned with the manually annotated samples.

Model	Accuracy
SVM-R_0	0.5486
SVM-Q_0R_0	0.5603
SVM-EU	0.5616
UM-R_0	0.5495
UM-Q_0R_0	0.5579
UM-EU	0.5638
UPM-Q_0R_0	0.5723
UPM-EU	0.5781
USM-EU	0.6022
RSM-EU	**0.6106**

Table 2: Accuracies of different models.

Firstly, the last computer's response R_0 is the basic feature that makes the prediction effective. Comparing the models involving Q_0-R_0 pair and EU with those only use R_0, we can easily find that the context provides more information about the trend of emotion.

SVM achieves comparable results with CNN based utterance model. We see that the convolutional process with a fixed-size filter encodes the similar information with n-gram features in SVM.

UPM outperforms UM with both Q_0-R_0 pair

[3] www.tensorflow.org/

and EU. UPM connects the abstract representation in the hidden layer, while UM connects the sentences into a whole as input. Although their structures are similar as shown in Figure 2, the logic depths of the two models are different. UM composes word embedding of all sentences in the convolutional process to learn an emotional representation, while UPM gets emotion features in two steps (composing word representation in convolutional layer and then adding mapped the sentence embedding after pooling). In practice, UPM works as a hierarchical compositional model. Such strategy makes the internal compositional process more flexible and expressive. In this way, the hidden layer simulates the relation between sentences in a more appropriate manner.

The CRNN based models (USM-EU and RSM-EU) achieve a significant improvement over other approaches including UPM (according to the two-sided paired t-test with a confidence level of $\alpha = 0.05$). UPM maps the representation of sentences to the same space and adds up the mapping result as the conversation representation. However, such process is less expressive than CRNN. In the test process, the weight matrix that mapping sentence representations in UPM is constant after training and the contribution of each sentence to the conversation is relatively fixed. While the GRUs in the CRNN could select the information resource flexibly through the reset gate r and update gate z, controlling the influence of particular sentence according to the context (Chung et al., 2014). Moreover, the gating process is a kind of multiplicative operation between sentence embeddings. Such multiplicative compositional functions are more expressive in simulating interaction between abstract features than additive ones (Socher et al., 2013; İrsoy and Cardie, 2015). Thus, CRNN based models handle the interaction between utterances in a more flexible way than UPM.

RSM is more effective than USM according to the results in Table 2. This is due to the fact that the gated operation makes it possible to adjust sentence representation according to R_0. Therefore, besides the interaction between adjacent utterances handled by the recurrent structure, the influence of interaction between R_0 and other utterances can be involved into the final representation and distance relation and consistency could be encoded.

4.5.2 Case Study

In order to illustrate the difference between USM and RSM in an intuitive way, we calculate the risk of negative feedback for each time step in these two recurrent models with the input of the session shown in Figure 1. The outputs of recurrent layer of each time step are used as inputs of the full-connection layer and the softmax regression results are considered as the probabilities of user's dissatisfaction.

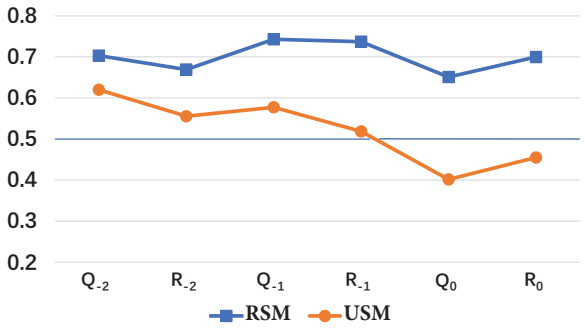

Figure 3: The probabilities of negative feedback for each utterance in sessions in Figure 1.

The line chart of the sequence of probabilities of the two models are shown in Figure 3. The tendencies of these two line are similar. It is due to the fact that the gates controlled by R_0 shrink the activations of recurrent layer and adjust the scales of values without changing the quadrant or feature space.

For the same utterance Q_{-2}, there's a difference between the probabilities of these two models, and the gap between the two line get more obvious after Q_{-1} is input. Both Q_{-2} and Q_{-1} contains negative emotions and possibly lead to dissatisfaction. Thus, the probabilities of negative feedback increase at the corresponding time step. However, the gated activations in RSM change more sharply. It indicates that the emotional inconsistency between R_0 and these two user messages lead to a further increasing of risk through the gated adjustment.

Finally, the RSM predicts that the negative feedback will occur in the next time step (with the probabilities larger than 0.5), which is true according to the corpus. However, the USM fails to make the correct prediction.

4.5.3 Comparison of Pre-training Strategies

As discussed in Subsection 4.2, different pre-training strategies are implemented during the

Models	NPT	UPT	BPT
UM-R_0	0.5365	0.5232	0.5495
UPM-Q_0R_0	0.5619	0.5547	0.5623
USM-EU	0.5865	**0.5808**	0.6022
RSM-EU	**0.5939**	0.5782	**0.6106**

Table 3: Comparison of accuracies with no pre-training (NPT), unbalanced pre-training (UPT) and balanced pre-training (BPT)

training process. The comparison of accuracy of these strategies are shown in Table 3.

Unbalanced pre-training strategy leads to a worse performance than only using manually labelled data. As discussed in 4.2.2, when a false rule is learned, a particular R_0 is associated with a wrong label, which hurts the performance obviously. Moreover, during the experimental process of unbalanced pre-training, it is observed that the models involving more context achieve better result than those only using R_0 as input. It is due to that the existence of the strong correlation between R_0 and the label itself is an inaccurate pattern, no matter whether the label is correct. The pre-training data encoding such strong correlation will makes the models ignore the context utterance and convergence to the local optimum only related to R_0.

However, the balanced pre-training dataset is effective in initializing the networks. The experimental results show that the balanced pre-training improves the performance of the networks. The underlying reason is that the pre-training process provides a better initialization for the networks, and the converging process of tuning continues based on an initial optimization.

4.5.4 Other Discussions

Directionality: Bi-directional and backward-directional recurrent networks are tested. Both structures lead to drop in accuracy (about 1%). We see that the last response R_0 is the essential determinant of emotion, the basic forward RNN structure has a bias on the last time steps for being free from the influence of small recurrent connection weight matrix. While adding backward-directional processing involving more parameters and weaken the influence of R_0 and Q_0-R_0 pair.

Filter Size: Inspired by the SVM baseline models performing not worse than the utterance model, we introduce 1×1 filters, working together with 2×1 ones, and such setting achieves an improvement by about 1% than only using 2×1 ones (from 0.6019 to 0.6106 with RSM). In practice, the larger filter size (e.g. 3, 4 or 5) leads to instability in the prediction. It is due to the fact that the utterances conversations are relatively short. Features within a uni-gram or bi-gram window is eligible for representing the emotional information. Although covering some sparse features, involving more larger filters results in risk of over-fitting.

5 Conclusion and Future Work

In this paper, we propose the problem of predicting users impending negative feedbacks by modelling the context queries and replies in human-agent conversation. Four kinds of influencing factors, (1) the computer's last response R_0, (2) the relation of last turn dialogue pair Q_0-R_0, (3) the sequence of all utterance and (4) the sequence of relation between utterances, are modelled with deep neural networks. The experimental results show that these factors indeed influence the emotional trend. We have encoded the possibility of dissatisfaction by representing the sequence of relation between utterances with a gated convolutional recurrent neural network. Tested on the real-world human-agent dialogue dataset, the proposed architecture outperforms the existing conversation models. Besides, balanced sampling on distance supervision labelled data are shown to be reliable in network pre-training.

The accuracies of prediction is only about 60%, we see that different users show different emotional feedbacks towards the same context. Thus, there are a few potential explorations: (1) build corpus on fine-grained emotional categories and (2) predict the emotional distribution on these categories instead of classifying into a certain one. Moreover, we would like to apply the emotional risk to the response ranking to improve the user experience of dialogue system.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This research is partially supported by National Natural Science Foundation of China (No.61672192, No.61572151, No.61602131) and the National High Technology Research and Development Program ("863" Program) of China (No.2015AA015405).

References

David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. *Luke, I am Your Father: Dealing with Out-of-Domain Requests by Using Movies Subtitles*.

Rafael E. Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *ACL 2012 System Demonstrations*, pages 37–42.

Junyoung Chung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Eprint Arxiv*.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2016. Language modeling with gated convolutional networks.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9:249–256.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. 2016. Classifying emotions in customer support dialogues in social media. In *Meeting of the Special Interest Group on Discourse and Dialogue*.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *Meeting of the Special Interest Group on Discourse and Dialogue*, pages 87–95.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *International Conference on Neural Information Processing Systems*, pages 2042–2050.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

Ozan İrsoy and Claire Cardie. 2015. Modeling compositionality with multiplicative recurrent neural networks. In *International Conference on Learning Representations (ICLR)*.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *Computer Science*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *Computer Science*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Eprint Arxiv*.

Linchuan Li, Zhiyong Wu, Mingxing Xu, Helen Meng, and Lianhong Cai. 2016. Combining cnn and blstm to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition. In *INTERSPEECH*, pages 1392–1396.

Weiyuan Li and Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications An International Journal*, 41(4):1742–1749.

Hsin Yih Lin, Changhua Yang, and Hsin Hsi Chen. 2007. What emotions do news articles trigger in their readers? In *SIGIR 2007: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands, July*, pages 733–734.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Acl-02 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *International Conference on Artificial Intelligence*, pages 1305–1311.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. *Computer Science*, (4).

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *Computer Science*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2009. Emotion classification using massive examples extracted from the web. In *COLING 2008, International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, Uk*, pages 881–888.

Ryoko Tokuhisa and Ryuta Terashima. 2006. Relationship between utterances and "enthusiasm" in non-task-oriented. In *Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–167.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *Computer Science*.

Mingxuan Wang, Zhengdong Lu, Hang Li, Wenbin Jiang, and Qun Liu. 2015. gencnn: A convolutional architecture for word sequence prediction. *Computer Science*.

Bowen Wu, Baoxun Wang, and Hui Xue. 2016. Ranking responses oriented to conversational relevance in chat-bots.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2016. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm.

Changhua Yang, Hsin Yih Lin, and Hsin Hsi Chen. 2007. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278.

Wenpeng Yin, Hinrich Schtze, Bing Xiang, and Bowen Zhou. 2015. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Computer Science*.

Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alex I. Rudnicky. 2016. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *Meeting of the Special Interest Group on Discourse and Dialogue*, pages 55–63.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A c-lstm neural network for text classification. *Computer Science*, 1(4):39–44.