FERNet: Fine-grained Extraction and Reasoning Network for Emotion Recognition in Dialogues

Yingmei Guo, Zhiyong Wu, Mingxing Xu
Department of Computer Science and Technology
Beijing National Research Center for Information Science and Technology
Tsinghua University, Beijing, China
guoym18@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn
xumx@tsinghua.edu.cn

Abstract

Unlike non-conversation scenes, emotion recognition in dialogues (ERD) poses more complicated challenges due to its interactive nature and intricate contextual information. All present methods model historical utterances without considering the content of the target utterance. However, different parts of a historical utterance may contribute differently to emotion inference of different target utterances. Therefore we propose Fine-grained Extraction and Reasoning Network (FERNet) to generate target-specific historical utterance representations. The reasoning module effectively handles both local and global sequential dependencies to reason over context, and updates target utterance representations to more informed vectors. Experiments on two benchmarks show that our method achieves competitive performance compared with previous methods.

1 Introduction

With the development of human-machine interaction (HMI) applications, textual dialogue scenes appear more frequently. These scenes request effective and high-performance emotion recognition systems helping in building empathetic machines (Young et al., 2018). Therefore, emotion recognition in dialogues (ERD) is getting growing attention from both academic and business community.

Different from non-conversation scenes, the ERD task poses a more complicated challenge of modeling context-sensitive dependencies. Most of existing approaches adopt Convolution Neural Network (CNN) (Krizhevsky et al., 2012), followed by a max-pooling layer to obtain utterance representations (Kim, 2014; Torres, 2018; Hazarika et al., 2018a,b; Majumder et al., 2019; Ghosal et al., 2019). The process proceeds without the guidance of the target utterance, thus generated historical utterance representations are indistinguishable toward different target utterances. Emotion recognition may fail in cases where historical utterances express various emotions toward various targets, which may confuse the emotion recognition of target utterances. As Figure 1 shows, for different target utterances B_1 and B_2, the model should attend the words "good service" and "bad food" in A_1, separately. In a word, it is desired to pay different attention to different words of a certain historical utterance to generate the target-specific historical utterance representation.

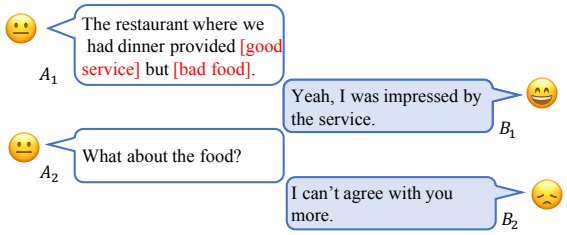

Figure 1: A dialogue shows that modeling intricate contextual information is crucial for emotion recognition.

In this paper, we propose Fine-grained Extraction and Reasoning Network (FERNet) to generate target-specific historical utterance representations conditioned on the content of target utterances by using the multi-head attention mechanism (Vaswani et al., 2017), extracting more fine-grained, relevant and contributing information for emotion recognition. Besides, we devise the reasoning module, which employs historical utterances as a sequence of triggers, and updates the representation of the target utterance to a more informed vector as it observes historical utterances through time. In the reasoning process, the module models both short-term and long-term sequential dependencies effectively. We demonstrate the effectiveness of our method on two benchmarks. Experimental results show that our method achieves competitive performance compared with previous methods.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 37–43

2 Related Work

Primitive approaches deal with the ERD task as simple solely-sentence emotion recognition task with no consideration of the historical information (Joulin et al., 2016; Chen et al., 2016; Yang et al., 2016; Chatterjee et al., 2019).

To exploit contextual information, Poria et al. (2017); Huang et al. (2019); Jiao et al. (2019); Hazarika et al. (2018a,b); Torres (2018) use RNN architecture, Hazarika et al. (2018b,a) use conversational memory networks (Sukhbaatar et al., 2015), Torres (2018); Jiao et al. (2019) use attention mechanism and Ghosal et al. (2019) uses graph neural network.

Besides, Majumder et al. (2019); Hazarika et al. (2018a) propose to keep track of states of individual speakers throughout the dialogue and Ghosal et al. (2019) incorporates speaker information into edge types.

Some of these works consider the context following the target utterance such as Luo et al. (2018); Saxena et al. (2018); Ghosal et al. (2019) and some variants of Majumder et al. (2019). However, this condition is quite incompatible with some practical situations like real-time dialogue systems in which we possess no future utterances while handling the target utterance. So in our paper, we only focus on the setting that only historical utterances can be utilized.

3 Proposed Model

Each dialogue D consists of two parts denoted as $D = \{(U, S)\}$, where $U = [u_1, u_2, ..., u_n]$ is a sequence of utterances ordered based on their temporal occurrence. $S = [s_1, s_2, ..., s_n]$ denotes corresponding speakers and n is the number of utterances in the dialogue. The ERD task aims to predict $Y = [y_1, y_2, ..., y_n]$, where $y_i \in C$ $(1 \leq i \leq n)$ denotes the underlying emotion of the utterance u_i. C is the set of candidate emotion categories. The FERNet consists of four successive modules: feature extraction module, attention module, reasoning module and output module. Figure 2 presents the overall architecture of the proposed model.

3.1 Feature Extraction Module

We use two multi-layer bidirectional Gated Recurrent Unit (bi-GRU) Networks (Tang et al., 2015) to accumulate contextual information from two directions for each word of target utterances and historical utterances, separately. The inputs consist of

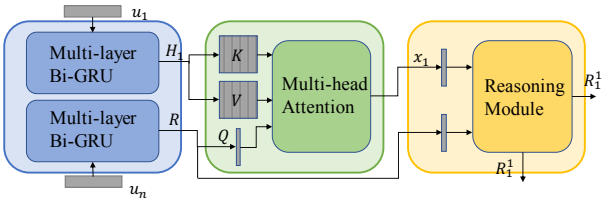

Figure 2: The overall architecture of the model.

300 dimensional pre-trained GloVe vectors (Pennington et al., 2014). The $k-$th contextual word representation $h_k^l = [\overrightarrow{h_k^l}; \overleftarrow{h_k^l}]$ is generated by concatenating the hidden states of the $k-$th time steps of forward and backward GRU, where l is the number of layers.

3.2 Attention Module

We utilize multi-head attention mechanism (Vaswani et al., 2017) to focus on more relevant parts of each historical utterance according to the target utterance. We also employ residual connection (He et al., 2016) followed by layer normalization (Ba et al., 2016) to make model training easier.

The target-specific representations of historical utterances are obtained by:

$$X = Concat(head_1, ...head_t)W^O \quad (1)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

where queries $Q = R$ are representations of target utterances, keys K and values V are contextual word representations for words of historical utterances. $W_i^Q \in \mathbb{R}^{d \times d_k}$,$W_i^K \in \mathbb{R}^{d \times d_k}$,$W_i^V \in \mathbb{R}^{d \times d_v}$ and $W^O \in \mathbb{R}^{td_v \times d}$ are parameter matrices, where d_k is the dimension of queries and keys, d_v is the dimension of values, d is the dimension of the output of feature extraction module and t is the number of heads. $X = [x_1, x_2, ..., x_{n-1}]$ are target-specific representations of historical utterances.

3.3 Reasoning Module

The reasoning module takes target-specific historical utterance representations $[x_1, x_2, ...x_{n-1}]$ and target utterance representations R as inputs. Target utterance representations are updated through time and layers.

Each unit in this module takes two inputs: R and $x_i (1 \leq i \leq n-1)$. The $t-$th unit updates R according to x_t by:

$$z_t = \alpha(x_t, R) = \sigma(W^z(x_t \circ R) + b^z) \quad (4)$$

$$r_t = \beta(x_t, R) = \sigma(W^r(x_t \circ R) + b^r) \quad (5)$$

$$\tilde{R}_t = \rho(x_t, R) = tanh(W^h[x_t; R] + b^h) \quad (6)$$

$$R_t = z_t r_t \tilde{R}_t + (1 - z_t) R_{t-1} \quad (7)$$

where z_t is the update gate, r_t is a reset function, $\tilde{R}_t$ is the candidate of updated representation of target utterance and R_t is the updated representation after observing the $t-$th historical utterance. σ is sigmoid activation, $tanh$ is hyperbolic tangent activation, $\circ$ is element-wise vector multiplication, and $[;]$ is vector concatenation along the last dimension. $W^z \in \mathbb{R}^{d\times d}$, $W^r \in \mathbb{R}^{d\times d}$, $W^h \in \mathbb{R}^{d\times 2d}$ are weight matrices, $b^z \in \mathbb{R}^d$, $b^r \in \mathbb{R}^d$, $b^h \in \mathbb{R}^d$ are bias terms.

Specifically, z_t measures the relevance between the target utterance representation and the tth historical utterance representation for fine-controlled gating. Compared with global attention computed over all historical utterances, the gate can be considered as local attention which models short-term sequential dependency. r_t is a reset function to determine how much previous information should be ignored by resetting the candidate of updated representation of target utterance.

As shown in Sukhbaatar et al. (2015), multi-hop can perform reasoning over multiple facts more effectively. So we stack several layers with outputs of the current layer used as inputs to the next layer. Besides, to model more abundant information, we compute $\overrightarrow{R_t^l}$ and $\overleftarrow{R_t^l}$ in both forward and backward directions and add them together to get R_t^l as the updated representation of the $t-$th unit in $l-$th layer:

$$R_t^l = \overrightarrow{R}_t^l + \overleftarrow{R}_t^l \quad (8)$$

Finally, we get the updated representation of target utterance $R^{update} = R_{n-1}^L$, where $n-1$ and L are the number of units and the number of layers in the reasoning module, respectively.

3.4 Output Module

After the feature extraction and reasoning modules, we obtain the updated representation of target utterance. To preserve original semantic content, we concatenate the updated representation and the original representation together:

$$R_{final} = [R^{update}; R] \quad (9)$$

We use a fully connected layer with softmax as activation to calculate emotion-class probabilities:

$$P = softmax(W^f R_{final} + b^f) \quad (10)$$

where $W^f \in \mathbb{R}^{d_{class}\times 2d}$ is a weight matrix, $b^f \in \mathbb{R}^{d_{class}}$ is a bias term and $P \in \mathbb{R}^{d_{class}}$ are emotion-class probabilities.

4 Experiment

Datasets We perform experiments on two benchmarks: IEMOCAP (Busso et al., 2008) and AVEC (Schuller et al., 2012). They are multimodal datasets involved in two-way dynamic conversations. In this paper, we only focus on using textual modality to recognize the emotion. The data distribution is shown in Appendices.

Evaluation Metrics We use accuracy (Acc.), F1-score (F1) and weighted average F1-socre (Average) as evaluation metrics for IEMOCAP dataset. Mean Absolute Error (MAE) and Pearson correlation coefficient (r) are used as metrics for AVEC dataset.

Baselines We compare the FERNet with following existing approaches: CNN (Kim, 2014), c-LSTM (Poria et al., 2017), c-LSTM+Attention (Poria et al., 2017), Memnet (Ba et al., 2016), CMN (Hazarika et al., 2018b), DialogueRNN (Majumder et al., 2019).

Training Details The training details such as hyper-parameters and settings we used are shown in Appendices.

4.1 Results

The overall results of experiments are shown in Table 1. We can see that our model outperforms baselines significantly on all evaluation metrics of both datasets. Specifically, our model surpasses DialogueRNN by 1.69% on weighted average F1-score. For AVEC dataset, our model lower mean absolute error by 0.03, 0.027, 0.009 and 0.31 for valence, arousal, expectancy and power, separately. We attribute the enhancement to the fundamental improvement of FERNet, which are generating target-specific representations of historical utterances and handling both short-term and long-term sequential dependencies.

4.2 Discussion and Analysis

Parameters We conduct experiments with different values of the number of historical utterances (N) and the number of layers of reasoning module (L) on the IEMOCAP dataset. Results are shown in Figure 4. We observe that as N increases, the performance of the model tends to be improved. This trend shows that adequate historical information

methods	IEMOCAP														AVEC							
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average		Valence		Arousal		Expectancy		Power	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	MAE	r	MAE	r	MAE	r	MAE	r
CNN	27.22	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18	0.545	-0.01	0.542	0.01	0.605	-0.01	8.71	0.19
c-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95	0.194	0.14	0.212	0.23	0.201	0.25	8.90	-0.04
c-LSTM+Attention	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19	0.189	0.16	0.213	0.25	0.190	0.24	8.67	0.10
Memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.2	59.00	55.72	55.10	0.202	0.16	0.211	0.24	0.216	0.23	8.97	0.05
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13	0.192	0.23	0.213	0.29	0.195	0.26	8.74	-0.02
DialogueRNN*	31.25	33.83	66.12	69.83	63.02	57.76	61.76	**62.50**	61.54	64.45	59.58	**59.46**	59.33	59.89	0.188	0.28	0.201	0.36	0.188	0.32	8.19	0.31
FERNet	38.89	**40.14**	72.65	**70.22**	67.19	**61.50**	66.47	62.43	68.90	**68.21**	50.39	58.63	**61.80**	**61.58**	**0.158**	0.44	**0.174**	0.43	**0.179**	0.37	**7.88**	0.36

Table 1: Performance of FERNet compared with baselines on the IEMOCAP dataset and AVEC dataset. Bold font denotes the best performances. * presents the state-of-the-art method in the setting that only historical utterances can be utilized.

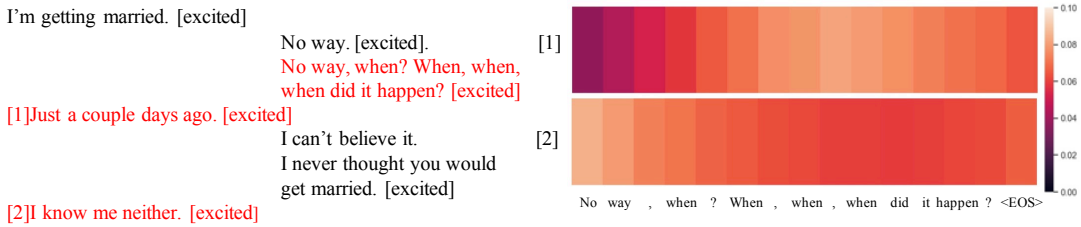

Figure 3: Average attention vectors across all attention heads for words of a historical utterance with regard to different target utterances. [1] shows the attention vector for the sentence "Just a couple days ago"; [2] shows the attention vector for the sentence "I know me either".

contributes to the performance of emotion recognition. However, a further increase of N degrades the performance of the model. It is mainly due to that there is too much-unrelated information confusing the model. As for L, the trend is similar to the parameter N. Models with hops in the range of 2-8 outperform the single layer variant. However, with L increasing, the reasoning module deepens and may cause the gradient vanishing problem which damages the performance of the model.

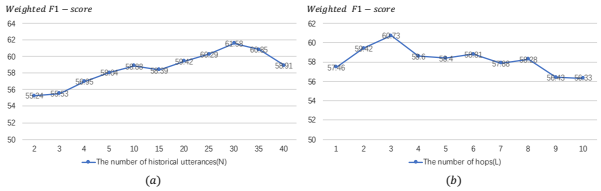

Figure 4: Performance of FERNet with different values of N and L. In (a), $L = 2$ and in (b), $N = 20$.

Ablation Study In order to demonstrate the effect of each module, we perform ablation studies. We compare the attention-based model with the attention-free model and replace the reasoning module with a memory network. As shown in Table 2, attention module and reasoning module both have a positive impact on model performance.

Case Study and Error Analysis We analyze the predicted results and find that misclassification often occurs when utterances are short. For example, our model classifies "what? " as "neutral", but the label is "excited". We think it is due to the lack of visual and audio modality. In this utterance, high pitched audio can provide vital information for recognizing the emotion. Besides, we find our model misclassifies several "excited" utterances as "happy" utterances, several "sad" utterances as "frustrated" utterances, and vice versa. The reason is that it is hard for the model to distinguish the subtle difference between these similar emotions.

methods	Acc.	F1
FERNet without attention	58.84	58.58
FERNet with memory network	59.77	59.33
FERNet	**61.80**	**61.58**

Table 2: Performance of variants of FERNet on the IEMOCAP dataset. Bold font denotes the best performances.

Besides, we perform qualitative visualization of the attention module. The dialogue in Figure 3 shows that for different target utterances, the model allocates different attention to words of a historical utterance. It demonstrates the effectiveness of the attention module.

5 Conclusion

In this paper, we propose FERNet to solve the ERD task. The model generates target-specific historical utterances according to the content of the target utterance using attention mechanism. The reasoning module effectively handles both local and global sequential dependencies to update the original representation of the target utterance to a more informed vector. Our model achieves competitive performance on two benchmarks.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. 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 2122–2132.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. 2019. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv:1904.00132*.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. *arXiv preprint arXiv:1904.04446*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Linkai Luo, Haiqing Yang, and Francis YL Chin. 2018. Emotionx-dlc: self-attentive bilstm for detecting sequential emotions in dialogue. *arXiv preprint arXiv:1806.07039*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883.

Rohit Saxena, Savita Bhat, and Niranjan Pedanekar. 2018. Emotionx-area66: Predicting emotions in dialogues using hierarchical attention network with sequence labeling. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 50–55.

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Johnny Torres. 2018. Emotionx-jtml: Detecting emotions with attention. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 56–60.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

A Appendices

Dataset	Partition	# of utterances	# of dialogues
IEMOCAP	train	5810	120
	test	1623	31
AVEC	train	4368	63
	test	1430	32

Table 3: Data distribution of IEMOCAP and AVEC datasets.

Training Details We use 10% of the training set as the validation set for hyper-parameters tuning. All tokens are lowercased with removal of stop words, symbols and digits, and sentences are zero-padded to the length of the longest sentence in the dataset. We alter the weight that each training instance carries when computing the loss to mitigate the influence of data imbalance. The weights are specific factors depending on corresponding emotions.

Hyper-parameters	**IEMOCAP**	**AVEC**
Optimizer	Adam	Adam
Learning rate	0.001	0.001
Batch size	16	16
Bi-GRU layer	2	2
Reasoning module layer	2	2
Historical utterance	30	20
GRU hidden size	150	150
Attention head	4	2
Attention hidden size	256	256

Table 4: Hyper-parameters and settings used for the two datasets.