Infusing Multi-Source Knowledge with Heterogeneous Graph Neural Network for Emotional Conversation Generation

Yunlong Liang[1][*], **Fandong Meng**[2], **Ying Zhang**[1], **Yufeng Chen**[1], **Jinan Xu**[1][†] **and Jie Zhou**[2]

[1]Beijing Key Lab of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{yunlongliang, zhying, chenyf, jaxu}@bjtu.edu.cn, {fandongmeng, withtomzhou}@tencent.com

Abstract

The success of emotional conversation systems depends on sufficient perception and appropriate expression of emotions. In a real-world conversation, we firstly instinctively perceive emotions from multi-source information, including the emotion flow of dialogue history, facial expressions, and personalities of speakers, and then express suitable emotions according to our personalities, but these multiple types of information are insufficiently exploited in emotional conversation fields. To address this issue, we propose a heterogeneous graph-based model for emotional conversation generation. Specifically, we design a *Heterogeneous Graph-Based Encoder* to represent the conversation content (*i.e.*, the dialogue history, its emotion flow, facial expressions, and speakers' personalities) with a heterogeneous graph neural network, and then predict suitable emotions for feedback. After that, we employ an *Emotion-Personality-Aware Decoder* to generate a response not only relevant to the conversation context but also with appropriate emotions, by taking the encoded graph representations, the predicted emotions from the encoder and the personality of the current speaker as inputs. Experimental results show that our model can effectively perceive emotions from multi-source knowledge and generate a satisfactory response, which significantly outperforms previous state-of-the-art models.

Introduction

Infusing emotions into conversation systems can substantially improve its usability and promote customers' satisfaction (Prendinger and Ishizuka 2005; Partala and Surakka 2004). Moreover, perceiving emotions sufficiently is the core premise of expressing emotions (Mayer and Salovey 1993). In real-life scenarios, humans can instinctively perceive complex or subtle emotions from multiple aspects, including the emotion flow of dialogue history, facial expressions and personalities of speakers, and then express suitable emotions for feedback. Figure 1 shows the organization of multi-source information in a dialogue graph and the relationship between them.
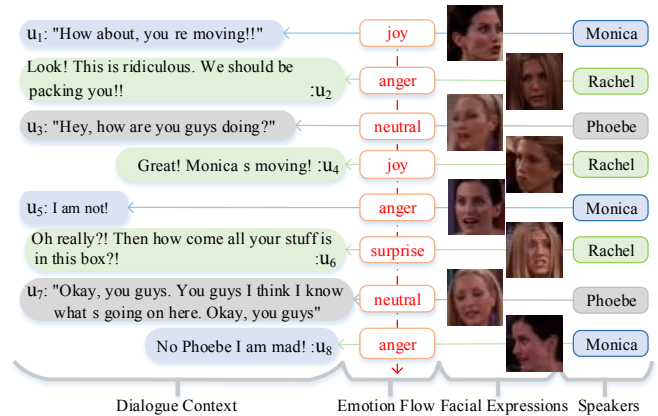

Figure 1: A dialogue example with multi-source knowledge (*i.e.*, the dialogue context, its emotion flow, facial expressions, and speakers). u_i: the i-th utterance.

Recently, there have been some efforts in emotional conversation systems, which can be roughly divided into two categories. (1) Several studies focus on building emotion-controllable conversation systems (Zhou et al. 2018a; Colombo et al. 2019; Song et al. 2019; Shen and Feng 2020) based on user-input emotion, which mainly apply various mechanisms to inject the specific emotion vector into the response generation process to enhance emotional expression. However, to some extent, it is a constrained scenario due to requiring an additional emotion input and these methods ignore the information from facial expressions and speakers' personalities (Madotto et al. 2019). (2) Others pay attention to automatically tracking the emotional state in dialogue history to generate emotional responses. Lubis et al.(2018) and Li et al.(2019) utilize an additional RNN-based encoder (Hochreiter and Schmidhuber 1997) to encode the discrete emotion label sequence (*i.e.*, emotion flow) and maintain it as the emotional context for the decoder, showing that the emotion flow can improve the emotion quality of responses. However, such specific emotion label sometimes can not fully express the complex and rich emotions of an utterance said by a speaker. For example, utterance u_3 in Figure 1 can be labeled as *neutral*, but its corresponding facial expression of the speaker *Phoebe* reflects an additional emotion of *joy*. And sometimes we can not correctly perceive the speaker's emotions only through textual expres-

sions, such as utterance u_5 in Figure 1, whose emotion (*i.e.*, *anger*) can be told by the corresponding facial expression of the speaker *Rachel*. Besides, the emotion is naturally inseparable from speaker's personality (Zhong et al. 2020). For instance, throughout the entire dialogue in Figure 1, utterances said by the speaker *Rachel* always have strong emotions due to her personality created in the *Friends* series, even though some of them contain no emotional words. Therefore, previous models perceive emotions only from the dialogue history and its emotion flow with discrete emotion labels while neglecting complex emotion flows from speakers' facial expressions and personalities may limit their performances.

In this paper, inspired by the capability of heterogeneous graph on capturing information from various types of nodes and relations (Zhang et al. 2019a), we propose a heterogeneous graph-based model to perceive emotions from different types of multi-source knowledge (*i.e.*, the dialogue history, its emotion flow, facial expressions, and speakers' personalities) and then generate a coherent and emotional response. Our model consists of a *Heterogeneous Graph-Based Encoder* and an *Emotion-Personality-Aware Decoder*. Specifically, we build a heterogeneous graph on the conversation content with four-source knowledge, and conduct representation learning over the constructed graph with the encoder to fully understand dialogue content, perceive emotions, and then predict suitable emotions as feedback. The decoder takes the graph-enhanced representation, the predicted emotions, and the current speaker's personality as inputs, and generates a coherent and emotional response.

We conduct experiments on the emotional conversation dataset MELD (Poria et al. 2019) to evaluate our model with both automatic and human evaluation. Experimental results show that comparing with competitive baselines (*e.g.*, the state-of-the-art model named ReCoSa (Zhang et al. 2019b)), our model can achieve sufficient emotion perception and generate more relevant responses with appropriate emotions, through infusing multi-source knowledge via the heterogeneous graph neural network. Furthermore, to evaluate the generalizability of our model, we conduct experiments on the DailyDialog (Li et al. 2017) dataset. Although this dataset only contains knowledge from two sources, *i.e.*, the dialogue history and its emotion flow, our model can still generate more satisfactory responses compared with baselines. Our contributions can be summarized as follows:

- We propose a novel heterogeneous graph-based framework for perceiving emotions from different types of multi-source knowledge to generate a coherent and emotional response. To the best of our knowledge, we are the first to introduce heterogeneous graph neural networks for emotional conversation.

- Experimental results on two datasets suggest that our model yields new state-of-the-art performances in emotional conversation generation, which also demonstrate the generalizability of our model, which can be easily adapted to different number of information sources[1].

[1]https://github.com/XL2248/HGNN

Our Approach

We elaborate our approach from Task Definition, Architecture and Training Objective.

Task Definition

Given the 5-tuples $< U, F, E, S, s_{N+1} >$ as inputs, where $U = \{u_1, \cdots, u_N\}$ is the dialogue history till round N, $F = \{f_1, \cdots, f_N\}$ is facial expression sequence, $E = \{e_1, \cdots, e_N\}$ is emotion sequence, $S = \{s_1, \cdots, s_N\}$ is speaker sequence, and s_{N+1} is the speaker for the next response, this task is to generate a target response $Y = \{y_1, \cdots, y_J\}$ with J words for the $(N+1)$-th round (actually the u_{N+1}), which is coherent with the content as well as with appropriate emotions. The dialogue history U contains N utterances, where the i-th utterance $u_i = \{x_1, \cdots, x_M\}$ is a sequence with M words. Corresponding to the i-th utterance u_i said by the speaker s_i, the facial expression f_i is a vector extracted by the OpenFace (Baltrusaitis et al. 2018) toolkit[2] from the dialogue video[3], and the emotion label e_i is one of seven emotion categories {*anger, disgust, fear, joy, sadness, surprise, neutral*}.

Architecture

As shown in Figure 2, our architecture contains two components *Heterogeneous Graph-Based Encoder* and *Emotion-Personality-Aware Decoder*, where the encoder aims to understand the content and perceive emotions from conversation context with multi-source information, and the decoder serves for generating a coherent as well as emotional response. Specifically, our encoder mainly consists of four parts: (1) *Graph Construction*, constructing the heterogeneous graph with multi-source knowledge; (2) *Graph Initialization*, initializing different kinds of nodes; (3) *Heterogeneous Graph Encoding*, perceiving emotions and representing the conversation context based on the constructed graph; and (4) *Emotion Predictor*, predicting suitable emotions with the graph-enhanced representations for feedback. And our decoder takes the graph-enhanced representation, the predicted emotions, and the current speaker's personality as inputs to generate an appropriate emotional response.

Heterogeneous Graph-Based Encoder
Graph Construction. As shown in the left part of Figure 2, we construct the heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{E}$ for the edges represents the relations between nodes. Specifically, we consider four types of heterogeneous nodes, where each node is an utterance in the dialogue history U, or a facial expression in F, or an emotion category in E or a speaker in S. Then, we build edges $\mathcal{E}$ among these nodes because there is a natural and close connection between two nodes: 1). between two utterances that are adjacent or said by the same speaker; 2). between an utterance and its corresponding facial expression; 3). between an utterance and its corresponding emotion; 4). between an utterance and its corresponding speaker

[2]https://github.com/TadasBaltrusaitis/OpenFace
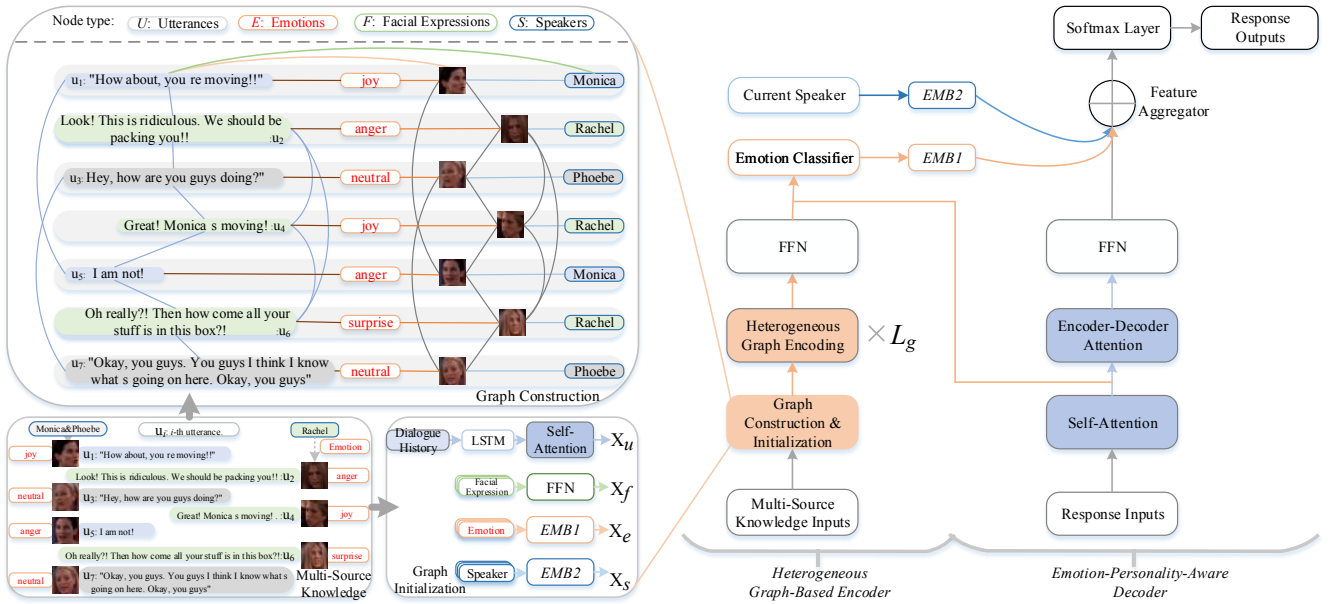[3]Each utterance is labeled with a video in MELD dataset.

Figure 2: The model architecture (on the right) together with an input example (on the left). The model consists of two parts: (1) *Heterogeneous Graph-Based Encoder* to encode multi-source knowledge via heterogeneous graph neural network and (2) *Emotion-Personality-Aware Decoder* to generate a suitable response according to the graph-enhanced representation, the emotions predicted by the encoder and the personality of the current speaker. Note that we only demonstrate the relations among utterance u_1, its emotion category, and its associated facial expression/speaker in the 'Graph Construction' while omitting those for dialogue rounds $u_2 \sim u_7$ for simplicity.

who said it; 5). between two facial expressions that are adjacent or from the same speaker; 6). between a facial expression and its corresponding speaker; 7). between a facial expression and an emotion that correspond to the same utterance.

Graph Initialization. We take one dialogue with N-round history as an example to describe how to initialize the four types of nodes.

For **utterance nodes**, we firstly use a Long Short-Term Memory (LSTM) network (Bahdanau, Cho, and Bengio 2015) to encode each utterance u_i and take the last hidden state of the LSTM as the primary representation $\mathbf{h}_{u_i}$ of the utterance node. By doing so, we obtain the primary representations of N utterance nodes $\{\mathbf{h}_{u_1}, \cdots, \mathbf{h}_{u_N}\}$. We next employ multi-head self-attention (Vaswani et al. 2017) to generate the contextual representation of each utterance node by aggregating the features from other utterances of the dialogue history. We also use the position embeddings ($\mathbf{PE}$) (Vaswani et al. 2017) to distinguish different positions among utterances by concatenating them to the primary utterance representations, *i.e.*, $\mathbf{H}_u = \{[\mathbf{h}_{u_1}; \mathbf{PE}_N], \cdots, [\mathbf{h}_{u_N}; \mathbf{PE}_1]\}$, where we set the position index start from the last history utterance to the first one and $[\cdot; \cdot]$ denotes concatenation operation. Finally, we calculate the contextual representations for the N utterance nodes as follows:

$$\mathbf{X}_u = \text{MultiHead}(\mathbf{H}_u, \mathbf{H}_u, \mathbf{H}_u), \ \mathbf{X}_u \in \mathbb{R}^{N \times d_u},$$

where $\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is a multi-head self-attention function, which takes a query matrix $\mathbf{Q}$, a key matrix $\mathbf{K}$, and a value matrix $\mathbf{V}$ as inputs, and d_u is the dimension of $\mathbf{X}_u$. We take $\mathbf{X}_u$ as the initialized feature of utterance nodes.

For **facial expression nodes**, we firstly extract the original facial expression features $\mathbf{G}_f = \{\mathbf{g}_{f_1}, \cdots, \mathbf{g}_{f_N}\}$, where $\mathbf{g}_{f_i} \in \mathbb{R}^{d_g}$. Then, we apply a position-wise feed forward network $\text{FFN}(*)$ (Vaswani et al. 2017) to project $\mathbf{G}_f$ to textual vector space $\mathbf{X}_f$ as follows:

$$\mathbf{X}_f = \text{FFN}(\mathbf{G}_f), \ \mathbf{X}_f \in \mathbb{R}^{N \times d_f},$$

where d_f is the dimension of facial expression node representation. We take $\mathbf{X}_f$ as the initialized representation of facial expression nodes.

For **emotion nodes** and **speaker nodes**, we maintain trainable parameter matrices $EMB1$ and $EMB2$, which are randomly initialized and can be learned. $EMB1 \in \mathbb{R}^{7 \times d_e}$ stores seven emotion label features and $EMB2 \in \mathbb{R}^{Z \times d_s}$ stores the feature of Z speakers' personalities, where d_e and d_s is the dimension of emotion node representation and speaker node representation, respectively. Therefore, the initialized features of emotion nodes and speaker nodes are retrieved from them, denoted as $\mathbf{X}_e$ and $\mathbf{X}_s$, respectively.

Heterogeneous Graph Encoding. Then we introduce the encoding part of the heterogeneous graph neural network (HGNN, (Zhang et al. 2019a)) on the constructed heterogeneous graph for capturing features from various types of nodes and relations. First, we introduce the conventional adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where $|\mathcal{V}|$ denotes the number of all nodes. The values in $\mathbf{A}$ are either 1 or 0, *e.g.*, $\mathbf{A}_{ij}=1$ denotes there is an edge between node i and node j. Then we define four type-wise adjacency matrix $\mathbf{A}_\tau \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where $\tau \in \{u, f, e, s\}$. Specifically, $\mathbf{A}_\tau$ is generated by $\mathbf{A}$ multiplying the corresponding masked matrix $\mathbf{M}^\tau \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where $\mathbf{M}^\tau_{i,*}=1$ if the node i is of type

13345

τ, otherwise 0. We use $\mathbf{H}^l \in \mathbb{R}^{|\mathcal{V}| \times d}$ to denote hidden representations of nodes in the l-th layer[4].

Then, the HGNN considers various types of nodes and maps them into an implicit common space with their individual trainable matrices by:

$$\mathbf{H}^{l+1} = \sigma(\sum_{\tau \in \{u,f,e,s\}} \mathbf{A}_\tau \mathbf{H}^l \mathbf{W}^l_\tau + \mathbf{b}^l_\tau),$$

where σ is an activation function (*e.g.*, ReLU). $\mathbf{H}^{l+1}$ is obtained by aggregating features of their neighboring nodes $\mathbf{H}^l$ with different type node τ using its corresponding transformation matrix $\mathbf{W}^l_\tau \in \mathbb{R}^{d \times d}$ and bias $\mathbf{b}^l_\tau \in \mathbb{R}^d$. Initially, $\mathbf{H}^0 = [\mathbf{X}_u; \mathbf{X}_f; \mathbf{X}_e; \mathbf{X}_s]$. By stacking L_g such layers, HGNN can aggregate features from various nodes, and the $\mathbf{H}^{L_g}$ contains the representations of all nodes. Then, we get the final output $\mathbf{H}^{enc}$ by:

$$\mathbf{H}^{enc} = \mathrm{FFN}(\mathbf{H}^{L_g}), \ \mathbf{H}^{enc} \in \mathbb{R}^{|\mathcal{V}| \times d}.$$

If we ignore the heterogeneity of different node types, the HGNN will become a conventional homogeneous graph neural network (Kipf and Welling 2017):

$$\mathbf{H}^{l+1} = \sigma(\mathbf{A}\mathbf{H}^l \mathbf{W}^l + \mathbf{b}^l),$$

where $\mathbf{W}^l$ and $\mathbf{b}^l$ are layer-wise trainable weights independent of node types.

Emotion Predictor. After fully perceiving emotions from multiple different sources, the encoder stores the representation in the final layer. We transform the representation $\mathbf{H}^{enc}$ into a fixed-size vector through $\mathrm{Maxpooling}$. Then we apply a fully-connected layer to predict suitable emotions:

$$\mathbf{H}^{max} = \mathrm{Maxpooling}(\mathbf{H}^{enc}), \ \mathbf{H}^{max} \in \mathbb{R}^d,$$
$$\mathcal{P} = \mathrm{Softmax}(\mathbf{W}^p \mathbf{H}^{max}), \quad (1)$$

where $\mathbf{W}^p \in \mathbb{R}^{7 \times d}$ is trainable weight[5].

Emotion-Personality-Aware Decoder.
We aim to incorporate multi-sources knowledge with heterogeneous graph neural network for emotional perception, so we use a simple and general decoder, *i.e.* the self-attention based decoder (Vaswani et al. 2017), to generate the response word by word, as shown in the right part of Figure 2. Obviously, a more powerful emotion-aware decoder can be extended to our framework, such as (Zhou et al. 2018a).

For generating the t-th word y_t, we firstly take the previous words $y_{1:t-1}$ as inputs to get the representation with a multihead self-attention (future masked) as follows[6]:

$$\mathbf{H}^r = \mathrm{MultiHead}(\mathbf{R}, \mathbf{R}, \mathbf{R}),$$

where $\mathbf{R}$ is the embedding sequence of the already generated words of target response r, *i.e.*, $y_{1:t-1}$.

Then, we use another multi-head attention followed by an FFN layer, taking the response history representation $\mathbf{H}^r$ as query, and taking $\mathbf{H}^{enc}$ as key and value to output the representation $\mathbf{O} \in \mathbb{R}^{J \times d}$:

$$\mathbf{O} = \mathrm{FFN}(\mathrm{MultiHead}(\mathbf{H}^r, \mathbf{H}^{enc}, \mathbf{H}^{enc})).$$

To effectively incorporate the predicted emotions and the

---

[4]We set $d = d_u = d_f = d_e = d_s$.

[5]We omit the bias in Softmax layer, similarly hereinafter.

[6]We omit the layer normalization for simplicity, and you may refer to Vaswani et al. (2017) for more details.

speaker's personality into the generation process, we design a gate to dynamically control the contribution of these information:

$$\mathbf{O}^{es} = \mathbf{O} + g \odot \mathbf{E}_g + (1 - g) \odot \mathbf{S}_g,$$
$$g = \sigma([\mathbf{O}; \mathbf{E}_g; \mathbf{S}_g]\mathbf{W}^g + \mathbf{b}^g),$$
$$\mathbf{E}_p = \sum(\mathcal{P} \cdot EMB1), \ \mathbf{E}_p \in \mathbb{R}^d,$$

where $\mathbf{E}_p$ denotes the mixed emotional representation generated by the weighted sum of the emotion distribution $\mathcal{P}$ (predicted by the encoder as Eq. 1) and the emotion parameter matrix $EMB1$. The $\mathbf{S}_p$ is the feature of current speaker s_{N+1} retrieved from the speaker parameter matrix $EMB2$. We repeatedly concatenate $\mathbf{E}_p$ for J times, *i.e.*, $\mathbf{E}_g \in \mathbb{R}^{J \times d} = [\mathbf{E}_p; \cdots; \mathbf{E}_p]$, similarly, $\mathbf{S}_g \in \mathbb{R}^{J \times d} = [\mathbf{S}_p; \cdots; \mathbf{S}_p]$. $\mathbf{W}^g \in \mathbb{R}^{d \times d}$ and $\mathbf{b}^g \in \mathbb{R}^d$ are the trainable parameters.

Finally, a softmax layer is utilized to obtain the word probability by taking the emotion-personality-aware representation $\mathbf{O}^{es}$ as input. Therefore, the probability of word y_t is calculated as follows:

$$P(y_t|y_{1:t-1}; \mathcal{G}; \mathbf{E}_p; \mathbf{S}_p; \theta) = \mathrm{Softmax}(\mathbf{W}^o \mathbf{O}^{es}_t),$$

where $\mathcal{G}$ is the constructed graph; $\mathbf{W}^o \in \mathbb{R}^{|V| \times d}$ is trainable parameter, where V is the vocabulary size. The log-likelihood of the corresponding response sequence $Y = \{y_1, \cdots, y_J\}$ is:

$$P(Y|\mathcal{G}; \mathbf{E}_p; \mathbf{S}_p; \theta) = \prod_t P(y_t|y_{1:t-1}; \mathcal{G}; \mathbf{E}_p; \mathbf{S}_p; \theta).$$

Training Objective

Our model can be trained end to end, and the overall training objective consists of the response generation loss $\mathcal{L}_{MLL}$ and the emotion classification loss $\mathcal{L}_{CLS}$ as follows:

$$\mathcal{J} = (1 - \lambda)\mathcal{L}_{MLL} + \lambda\mathcal{L}_{CLS},$$
$$\mathcal{L}_{MLL} = \min(-\sum_{t=0}^J \log(y_t)),$$
$$\mathcal{L}_{CLS} = \min(-\log(\mathcal{P}[e_{N+1}])),$$

where λ is the discount coefficient to balance the generation loss and classification loss and e_{N+1} is the golden class label, $\mathcal{P}$ is calculated as in Eq. 1.

Experiments

Datasets

MELD (Poria et al. 2019). It is a multi-party and multi-modal emotional dialogue dataset from the Friends TV series, which contains textual, acoustic, video, and speakers information. We preprocess each video into a sequence of images with the speaker's facial expression according to the video resolution. Then we use the OpenFace toolkit (Baltrusaitis et al. 2018) to extract the facial features from each image and average these features of all images from one video as the final expression for the corresponding utterance. Each utterance in every dialogue is labeled with one of the seven emotion categories.

DailyDialog (Li et al. 2017). To verify the generalizability of our model across datasets, we conduct experiments

on DailyDialog, a larger scale dataset only containing textual utterances with the same emotion categories as MELD. Therefore, we remove the facial expression nodes and the speakers nodes from the constructed graph $\mathcal{G}$ to adapt our model to DailyDialog.

Comparison Models

Seq2Seq (Sutskever, Vinyals, and Le 2014): It is the simplest generation-based approach, which has been widely applied to generation tasks.

HRED (Serban et al. 2016): It is a hierarchical encoder-decoder network, which performs well due to its context-aware modeling ability.

E-HRED (Lubis et al. 2018): It is an extension of HRED by adding an additional encoder to represent the emotional labels of dialogue as the emotional context for the decoder.

ReCoSa (Zhang et al. 2019b): It uses self-attention to measure the relevance between response and dialogue history, and shows state-of-the-art performances on benchmark datasets. In order to make the comparison more convincing, we extend this model to utilize multi-source knowledge by concatenating other source features (*i.e.*, facial expressions, emotions, and speakers' personalities) to dialogue history representations, and train it with the same loss function $\mathcal{J}$. Besides, we replace its original decoder with *Emotion-Personality-Aware Decoder* for fair comparison.

GNN (Kipf and Welling 2017): We adapt our architecture to a conventional graph-based network by taking all nodes as homogeneous ones.

Settings and Hyperparameters

For a fair comparison, we train our models with the same settings as comparison models. Specifically, the word embedding dimension and hidden size are set to 128 and 256, respectively. The d_u, d_f, d_e, d_s and d are set to 256. The number of encoder layers L_g is 2. The number of attention heads of ReCoSa and our model are set to 4. The discount coefficient λ is empirically set to 0.5. The dropout rate is set to 0.1 and the batch size is 16. Adam (Kingma and Ba 2014) is utilized for optimization with the learning rate 0.001. The trainable parameter matrices $EMB1$ and $EMB2$ are randomly initialized by Xavier (Glorot and Bengio 2010). The max turn of dialogue is 35 and the max sentence length is 50. We set the number of speakers Z to 13. Since most speakers just said several utterances in the whole corpus, to learn better representations of speakers' personality, we remove the speakers whose utterances are less than 30 pieces and finally retain 12 speakers. To adapt a new speaker, we add UNK for unknown speakers and there are total 13 speakers. For the MELD and DailyDialog, we follow ReCoSa (Zhang et al. 2019b) and set the vocabulary size (*i.e.*, V) to 5,253 and 17,657, respectively.

Evaluation Metrics

Automatic Metrics. We use PPL and BLEU scores (Xing et al. 2017) to evaluate the fluency and relevance of generated responses, respectively. We use D-1 and D-2 (Li et al. 2016c) to evaluate the degree of diversity. We use three embedding-based metrics (Liu et al. 2016)(average, greedy and extreme) to evaluate the semantic-level similarity between the generated responses and the ground truth. They are all widely applied in response generation (Tian et al. 2017; Chen et al. 2018; Xing et al. 2018). To measure the accuracy of emotion expression for generated responses, we follow (Zhou et al. 2018a; Song et al. 2019) and fine-tune a BERT-Based (Devlin et al. 2019) emotion classifier on the text utterances of MELD as one evaluation tool. The 'W-avg.' result of the evaluation tool is 62.55%, which is better than existing best classifier DialogueGCN (58.10%, also trained on MELD by Ghosal et al.(2019)). The weighted-average F1-score (Ghosal et al. 2019) (denoted as W-avg. for short) is the agreement between the predicted label by the evaluation tool and the ground-truth.

Human Evaluation. Following (Zhang et al. 2019b; Zhong, Wang, and Miao 2019), we sample 100 dialogues from the test set and exploit three annotators to evaluate the content and emotion quality. We then randomly disorder the responses generated by each comparison model. For each sample, annotators are asked to score a response from content level and emotion level independently:

- +2: (content) The response is fluent, coherent and meaningful / (emotion) The response conveys accurate and appropriate emotion.

- +1: (content) The response is fluent but irrelevant / (emotion) The response conveys inaccurate but appropriate emotion for the dialogue history.

- 0: (content) The response is not fluent and irrelevant / (emotion) The response conveys inaccurate and inappropriate emotion.

Results and Analysis

Automatic Evaluation Results

From the results in Table 1, we see that incorporating additional knowledge indeed helps models to generate a response not only more relevant to the content but also with appropriate emotions, suggesting the necessity of infusing multi-source knowledge, especially by our HGNN. We can conclude that from Table 1:

(1) In all three settings ('One Source', 'Two Sources', and 'Four Sources'), Table 1 shows that our HGNN consistently surpasses other competitors on most evaluation metrics. This suggests that infusing multi-source knowledge with heterogeneous graph has a positive influence on response generation, and demonstrates the superiority of our model on understanding content and emotion, which yields new state-of-the-art performances.

2) In 'Two Sources' setting, our HGNN outperforms the best ReCoSa (Zhang et al. 2019b) on MELD by 2.47↓, 1.30↑, 1.41↑, 0.02↑, 0.01↑, 0.01↑ and 1.45↑ on PPL, D-1, D-2, Avg., Ext., Gre., and W-avg., respectively, showing that our HGNN indeed generates more fluent, diverse, coherent and emotional responses than baselines. Although our HGNN achieves comparable BLEU scores in contrast to ReCoSa (*i.e.* 1.91 vs. 2.00 on MELD, 0.75 vs. 0.72 on DailyDialog), the human evaluation will further verify the

	Methods	MELD								DailyDialog							
		PPL	BLEU	D-1	D-2	Avg.	Ext.	Gre.	W-avg.	PPL	BLEU	D-1	D-2	Avg.	Ext.	Gre.	W-avg.
One Source	Seq2Seq	108.29	1.34	.54	2.13	.544	.365	.385	28.08	85.29	.20	.013	.085	.579	.344	.394	65.40
	HRED	121.51	1.77	.086	1.55	.566	.362	.410	28.56	82.02	.27	.015	.20	.577	.322	.399	69.49
	ReCoSa	114.19	2.11	1.38	2.57	.565	.370	.413	28.96	80.43	.74	.17	.23	.584	.353	.390	68.76
Two Sources	E-HRED	116.14	1.87	1.23	3.77	.566	.369	.412	29.67	81.48	.54	.12	.23	.598	.350	.406	70.49
	ReCoSa	108.51	2.00	1.22	4.54	.568	.365	.419	30.13	80.28	.72	.13	.24	.602	.349	.435	69.63
	GNN (Ours)	108.47	2.06	1.75	2.58	.571	.364	.419	30.18	78.81	.65	**.24**	.40	.617	.372	.431	69.93
	HGNN (Ours)	106.04	1.91	2.52	5.95	.583	.376	.427	31.58	**76.59**	**.75**	.22	**.47**	**.641**	**.418**	**.510**	**71.16**
Four Sources	ReCoSa	102.37	2.49	2.07	6.88	.571	.364	.421	30.66	-	-	-	-	-	-	-	-
	GNN (Ours)	100.66	2.15	2.92	4.63	.578	.372	.426	31.25	-	-	-	-	-	-	-	-
	HGNN (Ours)	**96.90**	**2.80**	**4.84**	**9.15**	**.593**	**.394**	**.437**	**34.49**	-	-	-	-	-	-	-	-

Table 1: Automatic evaluation results of generated responses on two test sets. Three settings mean different knowledge is used, namely 'One Source' for only dialogue history, 'Two Sources' for dialogue history and its emotion flow, and 'Four Sources' for dialogue history, its emotion flow, facial expressions, and speakers' personalities. Note: 'Avg.', 'Ext.', and 'Gre.' denote three embedding-based metrics (average, greedy, and extreme, respectively). 'W-avg.' denotes weighted-average F1-score of seven emotion categories, which is the agreement between the predicted label by the BERT-Based evaluation tool and the ground-truth. Lower PPL is better and higher is better for other metrics.

	Methods	MELD				DailyDialog			
		+2	+1	0	Kap.	+2	+1	0	Kap.
One.	Seq2Seq	27.0	43.3	39.7	.531	25.3	51.0	23.7	.562
	HRED	27.3	31.7	41.0	.547	25.7	51.6	22.7	.504
	ReCoSa	27.3	35.4	37.3	.472	27.0	51.0	22.0	.564
Two.	E-HRED	27.7	34.0	38.3	.420	27.3	52.4	20.3	.478
	ReCoSa	28.7	41.0	30.3	.459	29.0	51.0	20.0	.495
	GNN	29.0	43.7	27.3	.499	32.3	45.0	22.7	.474
	HGNN	29.7	44.3	26.0	.467	**34.3**	43.7	22.0	.514
Four.	ReCoSa	34.6	39.7	25.7	.482	-	-	-	
	GNN	35.3	44.4	20.3	.518	-	-	-	
	HGNN	**36.0**	42.0	22.0	.505	-	-	-	

Table 2: Human evaluation on content quality. 'Kap.': Kappa.

	Methods	MELD				DailyDialog			
		+2	+1	0	Kap.	+2	+1	0	Kap.
One.	Seq2Seq	13.0	41.3	45.7	.637	10.7	47.3	42.0	.608
	HRED	13.7	45.7	40.6	.548	10.7	48.0	41.3	.494
	ReCoSa	14.3	45.0	40.7	.638	12.0	47.3	40.7	.577
Two.	E-HRED	15.3	40.4	44.3	.649	12.3	48.7	39.0	.593
	ReCoSa	15.3	48.7	36.0	.687	15.0	41.3	43.7	.602
	GNN	15.3	44.4	4.3	.568	15.3	46.4	38.3	.567
	HGNN	17.7	41.3	41.0	.593	**18.0**	45.0	37.0	.627
Four.	ReCoSa	19.7	50.6	29.7	.604	-	-	-	-
	GNN	22.7	49.0	28.3	.643	-	-	-	-
	HGNN	**24.7**	41.0	34.3	.618	-	-	-	-

Table 3: Human evaluation on emotion quality.

Methods	MELD				DailyDialog			
	E.	R.	G.	H.	E.	R.	G.	H.
E.	-	40.6	34.7	28.4	-	39.4	36.5	31.1
R.	59.4	-	39.3	35.7	60.6	-	42.7	35.0
G.	65.3	60.7	-	38.5	63.5	57.34	-	38.4
H.	**71.6**	64.3	61.5	-	**68.9**	65.00	61.6	-

Table 4: Preference test (%) between any two models in 'Two-Sources' setting. 'E.': E-HRED. 'R.': ReCoSa. 'G.': GNN. 'H.': HGNN.

effectiveness of our model on generating relevant responses in next section. Furthermore, we validate our HGNN on a larger dataset DailyDialog and the results show the generalizability of our model.

(3) Compared with GNN under the same settings, our HGNN always surpasses GNN by large margins on both datasets, especially in 'Four Sources' setting (*i.e.*, 3.76↓ on PPL, 0.65↑ on BLEU, 1.92↑ on D-1, 4.52↑ on D-2, 0.015↑ on Avg., 0.02↑ on Ext., 0.01↑ on Gre., and 3.24↑ on W-avg.), demonstrating the effectiveness and superiority of the heterogeneous graph on dealing with the various types of multi-source knowledge.

Human Evaluation Results

Results of content quality and emotion quality are shown in Table 2 and 3. The Fleiss' kappa (Fleiss and Cohen 1973) for measuring inter-rater agreement is included as well. All models have "moderate agreement" or "substantial agreement". For the MELD dataset, our HGNN have more +2 ratings than comparison models on both content-level and emotion-level (t-test (Koehn 2004), *p*-value < 0.05) in different sources settings, showing that our model is better at

generating coherent as well as emotional responses than others. In 'Four Sources' setting, the HGNN model achieves further improvements on +2 metric, which is also far beyond other baselines. This demonstrates that our model can effectively infuse facial expressions and speakers' personalities into the dialogue graph and generate accurate emotions. For the DailyDialog dataset, our HGNN model also obtains very obvious promotion on +2 metric, suggesting the generalizability of our model, which is consistent with the automatic evaluation. Furthermore, we conduct preference test in Table 4. We can see that HGNN is significantly preferred against other models (t-test, p-value < 0.05) under

#	Models	PPL	BLEU	D-1	D-2	Avg.	Ext.	Gre.	W-avg.
1	HGNN	**96.90**	**2.80**	**4.84**	**9.15**	**.593**	**.394**	**.437**	**34.49**
2	- image	113.94	2.17	2.31	5.95	.568	.367	.413	31.54
3	- emotion	111.42	2.09	2.18	6.37	.571	.364	.419	30.18
4	- all speakers	100.95	2.50	2.08	8.60	.573	.371	.418	32.92
5	- cur. speaker	105.95	2.62	2.15	8.87	.577	.378	.421	33.24

Table 5: Ablation study on MELD. '- cur. speaker' means removed it from the decoder while maintaining it in HGNN.

'Two Sources' scenario. Obviously, the coherent and emotional responses generated by our HGNN are more attractive to users than the responses produced by baselines.

Analysis

Ablation Study. Table 5 shows the results of the ablation study about heterogeneous nodes in HGNN. Row 1 is the results of our full model, which incorporates not only the dialog history but also its emotion flow, the facial expressions (*i.e.*, image), and the speakers. Rows 2∼4 are the results of removing the information of corresponding nodes. From the results, we can conclude that after removing each heterogeneous node information, the response quality of content and emotion is differently lowered, suggesting that various information of heterogeneous nodes is crucial for understanding content, and emotional perception and expression (row 1 vs. rows 2∼4), especially for emotion flow (row 3).

***Emotion Predictor* of the Encoder.** We also evaluate the performance of the *Emotion Predictor* from the *Heterogeneous Graph-Based Encoder* to show whether the encoder can fully perceive emotions from multi-source knowledge and predict suitable emotions. The results shown in Table 6 suggest that our encoder can more effectively perceive emotions from multi-source knowledge than baselines, and therefore predict more accurate emotions for the decoder.

Case Analysis. To give an insight on whether the emotion of the generated response is expressed appropriately, we provide some examples from MELD test dataset in Figure 4. It shows that HGNN can generate coherent responses with suitable emotions. We can see that the methods in 'One Source' setting generate some neutral and safe responses (*e.g.*, "uh-huh. I'm sorry." in example 1 or "I don't know!" in example 2 by Seq2Seq) due to lacking of additional emotion perception source. Although these methods in 'Two Sources' can generate emotional responses (*e.g.*, "Yeah? I'm sorry, I'll talk to you later." by E-HRED in example 1 or "Yeah, I'm pretty sure I'm still gonna find that" by ReCoSa in example 2, may contain *anger* and *neutral* emotion, respectively.), the emotions of the responses are not suitable with the dialogue history. In 'Four Sources' setting, although the ReCoSa and GNN generate more relevant responses after incorporating the facial expression and speakers' personalities, they can not express an appropriate emotion with the dialogue history. And the HGNN generates a coherent as well as emotional response, suggesting that sufficiently understanding content and perception of emotion via heterogeneous graph from various types of knowledge can effectively

Figure 3: A dialogue example with different speaker inputs for measuring the speaker's personalities.

	Methods	MELD	DailyDialog
	E-HRED	31.03	79.67
Two Sources	ReCoSa	32.19	81.60
	GNN (Ours)	32.35	81.84
	HGNN (Ours)	33.52	**82.70**
	ReCoSa	33.41	-
Four Sources	GNN (Ours)	34.49	-
	HGNN (Ours)	**36.28**	-

Table 6: The 'W-avg.' F1-score (%) for the *Emotion Predictor* of different methods on test sets.

enhance the quality of responses. Besides, we also provide one example to show speaker's personalities in Figure 3. Under the same setting, we can observe that the generated responses given different speakers are usually personalized (*i.e.* speaker's personality) with the same or close meaning.

Related Work

Dialog Systems. In the past few years, previous studies mainly focus on improving the content quality of dialogue (Li et al. 2016a; Zhang et al. 2019b), and only a little work pays attention to improving the emotion quality of dialogue (Zhou et al. 2018b; Rashkin et al. 2019). Researchers firstly set user-input emotions directly for response generation (1), such as emotion-controllable conversation systems, which contain some representative work (Zhou et al. 2018a; Colombo et al. 2019; Zhou and Wang 2018; Huang et al. 2018; Sun et al. 2018). To automatically learn emotions (2), researchers track the emotion flow of dialog history (Lubis et al. 2018; Lin et al. 2019; Li et al. 2019, 2020; Majumder et al. 2020) and generate emotion-rich responses (Zhong, Wang, and Miao 2019; Asghar et al. 2018), where they incorporate Valence, Arousal, and Dominance embeddings (Warriner, Kuperman, and Brysbaert 2013; Mohammad 2018) into their models to provide additional affective knowledge. Besides, (3) multi-modal studies also have attracted much attentions in conversation systems such as video-grounded dialogue system (Yoshino et al. 2019) and

		(a) Dialogue example.				(b) Dialogue example.	
Dialogue History	Ross	u_1: hey ! oh listen , i was just clearing some space for your stuff	joy	Chandler	u_1: Someone on the subway licked my neck! Licked my neck!!	disgust	
	Rachel	u_2: oh thanks , but listen , i was just at monica's and she and chandler had a big fight and they're not moving in	sadness	Phoebe	u_2:Oh Willie's alive!	surprise	
	Ross	u_3: what do you mean , they're not moving in ? they-they're still moving in right ?	surprise	Chandler	u_3: What are you guys doing?	neutral	
	Rachel	u_4:no-no , they just had a big blowout over what to do with my room	neutral	Monica	u_4:Oh, my mom called, they're gonna run our engagement announcement in the local paper, so we're looking for a good picture of us.	joy	
	Ross	u_5: what ? ! over a stupid room	surprise	Chandler	u_5: Oooh, I'm afraid that does not exist.	neutral	
	Rachel	u_6: yeah , i feel kinda bad for them , but i'm also really psyched cause i do n't have to move in here !	joy	Monica	u_6:That's not true, there are great pictures of us!	surprise	
	Ross	u_7:		Chandler	u_7: No, there are great pictures of you standing next to a guy who's going like this	neural	
				Phoebe	u_8:		
	Golden	u_7: oh no , yeah no , that part's great !	joy	Golden	u_8: Oh, my God! That's the creep that you're with at the	surprise	
One Source	Seq2Seq	uh-huh. I'm sorry.		Seq2Seq	I do n't know!		
	HRED	so whose is it ?		HRED	Well I don't care how hot it is it's uh, it's wrong		
	ReCoSa	well, I just wanted to take her to the museum like you from, and the baby, right now.		ReCoSa	Okay, well y'know what , there's gonna be one		
Two Sources	Emo-HRED	Yeah? I'm sorry, I'll talk to you later .		Emo-HRED	What?! , y'know what? .		
	ReCoSa	There's only room for one day, and I ca n't stop it's about there.		ReCoSa	Yeah, I'm pretty sure I'm still gonna find that		
	GNN (Ours)	Y'know it was kinda stupid.		GNN (Ours)	Oh, my God! well I call it ! !		
	HGNN (Ours)	What? something wrong?		HGNN (Ours)	Y'know, I do n't think it's cool.		
Four Sources	ReCoSa	I ca n't believe we have to make it, why have money soon who were gonna be in the skin.		ReCoSa	I know, I'm the one of the last one who needs to is looking at our first time I do n't even , he thinks when she' s a charming guy.		
	GNN (Ours)	Oh, I did n't know you would say that.		GNN (Ours)	Okay, you guys. you're not fine		
	HGNN (Ours)	Oh, no, no, I just, I think it's good for you moving in.		HGNN (Ours)	Wow!! I find a picture looking good!		

Figure 4: Sample responses generated by all models.

visual question-answering (Antol et al. 2015), aiming to answer human queries grounded a video or image.

The differences between our model and those in (1) and (2) are: they only perceive emotions from text and ignore other source knowledge, *e.g.*, the facial expressions and speakers' personalities, where the speaker information has been shown helpful for conversation systems (Li et al. 2016b). The deep difference from (1) is: we focus on automatically learning emotions and then expressing it rather than emotion-controllable conversation systems (*i.e.* a limited scenario for expressing a user-input emotion). Furthermore, we expect to construct a practical emotional conversation system, *i.e.* infusing multi-source knowledge with heterogeneous graph neural network for improving emotional conversation, so we employ a simple and general decoder. Obviously, a powerful emotion-aware decoder can be integrated into our framework, such as (Zhou et al. 2018a; Colombo et al. 2019; Song et al. 2019; Shen and Feng 2020), which will be considered in our future work to further improve the emotional performance. In contrast to tasks in (3), we mainly utilize multi-modal information for emotion perception and expression in response generation.

Heterogeneous Graph for NLP. Heterogeneous graph neural networks (Shi et al. 2017; Sun et al. 2009) can deal

with various types of nodes and edges and have more advantages than homogeneous graph neural networks (Wang et al. 2019; Zhang, Cui, and Zhu 2020). Its superiority has been verified in many natural language processing (NLP) tasks, such as graph representation learning (Hong et al. 2020), reading comprehension (Tu et al. 2019), text classification (Linmei et al. 2019; Liang et al. 2019, 2020) and extractive document summarization (Wang et al. 2020). Inspired by the success of heterogeneous graph neural network, we first introduce it to emotional conversation generation to gain a better understanding of content and fully perceive emotions from multi-source knowledge, and then produce a satisfactory response.

Conclusion and Future Work

We propose a heterogeneous graph-based framework to understand dialogue content and fully perceive complex and subtle emotions from multi-source knowledge to generate coherent and emotional response. Experimental results and analysis demonstrate the effectiveness and generalizability of our model, which can be easily adapted to different number of knowledge sources. In the future, we would like to infuse knowledge from more sources and further investigate various relations between them to further improve the quality of responses.

Acknowledgements

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*, 2425–2433.

Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; and Mou, L. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, 154–166. Springer.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. IEEE.

Chen, H.; Ren, Z.; Tang, J.; Zhao, Y. E.; and Yin, D. 2018. Hierarchical Variational Memory Network for Dialogue Generation. In *WWW*, 1653–1662.

Colombo, P.; Witon, W.; Modi, A.; Kennedy, J.; and Kapadia, M. 2019. Affect-Driven Dialog Generation. In *NAACL*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Fleiss, J. L.; and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33(3).

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *EMNLP-IJCNLP*.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W.; and Titterington, M., eds., *Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Machine Learning Research*, 249–256. Chia Laguna Resort, Sardinia, Italy: PMLR. URL http://proceedings.mlr.press/v9/glorot10a.html.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hong, H.; Guo, H.; Lin, Y.; Yang, X.; Li, Z.; and Ye, J. 2020. An Attention-based Graph Neural Network for Heterogeneous Structural Learning. In *AAAI*.

Huang, C.; Zaïane, O.; Trabelsi, A.; and Dziri, N. 2018. Automatic Dialogue Generation with Expressed Emotions. In *NAACL*.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980. URL http://arxiv.org/abs/1412.6980.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Koehn, P. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP*, 388–395. URL https://www.aclweb.org/anthology/W04-3250.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL*.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A Persona-Based Neural Conversation Model. In *ACL*.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016c. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*.

Li, Q.; Chen, H.; Ren, Z.; Chen, Z.; Tu, Z.; and Ma, J. 2019. EmpGAN: Multi-resolution Interactive Empathetic Dialogue Generation. *arXiv preprint arXiv:1911.08698* .

Li, Q.; Li, P.; Chen, Z.; and Ren, Z. 2020. Empathetic Dialogue Generation via Knowledge Enhancing and Emotion Dependency Modeling. *arXiv preprint arXiv:2009.09708* .

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*.

Liang, Y.; Meng, F.; Zhang, J.; Xu, J.; Chen, Y.; and Zhou, J. 2019. A Novel Aspect-Guided Deep Transition Model for Aspect Based Sentiment Analysis. In *EMNLP-IJCNLP*, 5572–5584.

Liang, Y.; Meng, F.; Zhang, J.; Xu, J.; Chen, Y.; and Zhou, J. 2020. A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. *CoRR* abs/2004.01951. URL https://arxiv.org/abs/2004.01951.

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 2019. MoEL: Mixture of Empathetic Listeners. In *EMNLP-IJCNLP*, 121–132.

Linmei, H.; Yang, T.; Shi, C.; Ji, H.; and Li, X. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *EMNLP-IJCNLP*, 4823–4832.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*.

Lubis, N.; Sakti, S.; Yoshino, K.; and Nakamura, S. 2018. Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach. In *AAAI*.

Madotto, A.; Lin, Z.; Wu, C.-S.; and Fung, P. 2019. Personalizing Dialogue Agents via Meta-Learning. In *ACL*.

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *EMNLP*, 8968–8979.

Mayer, J. D.; and Salovey, P. 1993. The intelligence of emotional intelligence. *Elsevier* 433–442.

Mohammad, S. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *ACL*.

Partala, T.; and Surakka, V. 2004. The effects of affective interventions in human–computer interaction. *Interacting with Computers* 16: 295–309. doi:10.1016/j.intcom.2003. 12.001.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*.

Prendinger, H.; and Ishizuka, M. 2005. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence* 19(3-4): 267–285. doi:10.1080/08839510590910174.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Shen, L.; and Feng, Y. 2020. CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation. In *ACL*.

Shi, C.; Li, Y.; Zhang, J.; Sun, Y.; and Yu, P. S. 2017. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29(1): 17–37.

Song, Z.; Zheng, X.; Liu, L.; Xu, M.; and Huang, X. 2019. Generating Responses with a Specific Emotion in Dialog. In *ACL*.

Sun, X.; Chen, X.; Pei, Z.; and Ren, F. 2018. Emotional human machine conversation generation based on seqgan. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 1–6. IEEE.

Sun, Y.; Han, J.; Zhao, P.; Yin, Z.; Cheng, H.; and Wu, T. 2009. RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. In *EDBT*, 565–576.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*, 3104–3112.

Tian, Z.; Yan, R.; Mou, L.; Song, Y.; Feng, Y.; and Zhao, D. 2017. How to Make Context More Useful? An Empirical Study on Context-Aware Neural Conversational Models. In *ACL*, 231–236.

Tu, M.; Wang, G.; Huang, J.; Tang, Y.; He, X.; and Zhou, B. 2019. Multi-hop Reading Comprehension across Multiple Documents by Reasoning over Heterogeneous Graphs. In *ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, D.; Liu, P.; Zheng, Y.; Qiu, X.; and Huang, X. 2020. Heterogeneous Graph Neural Networks for Extractive Document Summarization. In *ACL*, 6209–6219.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous Graph Attention Network. In *WWW*, 2022–2032.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45(4): 1191–1207.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic aware neural response generation. In *AAAI*.

Xing, C.; Wu, Y.; Wu, W.; Huang, Y.; and Zhou, M. 2018. Hierarchical recurrent attention network for response generation. In *AAAI*.

Yoshino, K.; Hori, C.; Perez, J.; D'Haro, L. F.; Polymenakos, L.; Gunasekara, R. C.; Lasecki, W. S.; Kummerfeld, J. K.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B.; Gao, X.; AlAmri, H.; Marks, T. K.; Parikh, D.; and Batra, D. 2019. Dialog System Technology Challenge 7. *CoRR* abs/1901.03461. URL http://arxiv.org/abs/1901.03461.

Zhang, C.; Song, D.; Huang, C.; Swami, A.; and Chawla, N. V. 2019a. Heterogeneous Graph Neural Network. In *SIGKDD*, 793–803.

Zhang, H.; Lan, Y.; Pang, L.; Guo, J.; and Cheng, X. 2019b. ReCoSa: Detecting the Relevant Contexts with Self-Attention for Multi-turn Dialogue Generation. In *ACL*.

Zhang, Z.; Cui, P.; and Zhu, W. 2020. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering* .

Zhong, P.; Wang, D.; and Miao, C. 2019. An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss. In *AAAI*, 7492–7500.

Zhong, P.; Zhang, C.; Wang, H.; Liu, Y.; and Miao, C. 2020. Towards Persona-Based Empathetic Conversational Models. In *EMNLP*.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.

Zhou, L.; Gao, J.; Li, D.; and Shum, H. 2018b. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *CoRR* abs/1812.08989. URL http://arxiv.org/abs/1812.08989.

Zhou, X.; and Wang, W. Y. 2018. MojiTalk: Generating Emotional Responses at Scale. In *ACL*.