Knowledge-Interactive Network with Sentiment Polarity Intensity-Aware Multi-Task Learning for Emotion Recognition in Conversations

Yunhe Xie*, **Kailai Yang***, **Chengjie Sun**†, **Bingquan Liu, Zhenzhou Ji**
Faculty of Computing, Harbin Institute of Technology, China
{xieyh, sunchengjie, liubq, jizhenzhou}@hit.edu.cn
klyang990203@gmail.com

Abstract

Emotion Recognition in Conversation (ERC) has gained much attention from the NLP community recently. Some models concentrate on leveraging commonsense knowledge or multi-task learning to help complicated emotional reasoning. However, these models neglect direct utterance-knowledge interaction. In addition, these models utilize emotion-indirect auxiliary tasks, which provide limited affective information for the ERC task. To address the above issues, we propose a Knowledge-Interactive Network with sentiment polarity intensity-aware multi-task learning, namely KI-Net, which leverages both commonsense knowledge and sentiment lexicon to augment semantic information. Specifically, we use a self-matching module for internal utterance-knowledge interaction. Considering correlations with the ERC task, a phrase-level Sentiment Polarity Intensity Prediction (SPIP) task is devised as an auxiliary task. Experiments show that all knowledge introduction, self-matching and SPIP modules improve the model performance respectively on three datasets. Moreover, our KI-Net model shows 1.04% performance improvement over the state-of-the-art model on the IEMOCAP dataset.

1 Introduction

Emotion recognition in conversation aims to identify each utterance's emotion from a conversation, which requires machines to understand the way of emotion expression during conversations (Poria et al., 2019b). Research on ERC helps in creating empathetic dialogue systems (Ghosh et al., 2017; Zhou et al., 2018), thus improving the overall human-computer interaction experience. Hence, the ERC task has a wide range of applications such as social media analysis (Li et al., 2019; Chatterjee et al., 2019) and intelligent systems like smart homes and chatbots (Young et al., 2018).

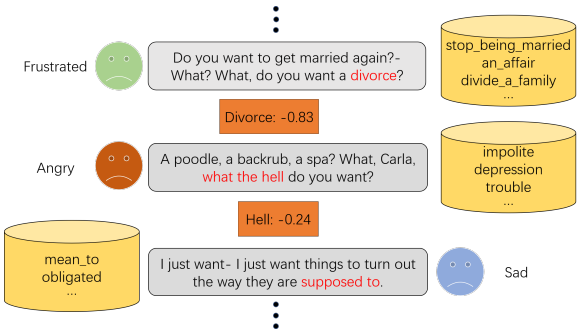

Figure 1: Illustration of a conversation where both sentiment lexicon and commonsense knowledge aid ERC task. Cylinders denote commonsense knowledge, and rectangles denote sentiment lexicon knowledge.

Unlike vanilla emotion recognition of sentences, context modeling for conversations is crucial for ERC models. Early Recurrent Neural Network (RNN)-based ERC works adopt memory networks to store historical conversation context (Hazarika et al., 2018b,a). Recent progress in Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019b) has benefitted many downstream tasks, such as dialogue systems (Henderson et al., 2020; Bao et al., 2020) and reading comprehension (Yang et al., 2019a; Shwartz et al., 2020). Nevertheless, Ilievski et al. (2021) indicate that PLMs lack some dimensions of knowledge, which may limit the performance of the corresponding downstream tasks. Hence most recent PLM-based ERC works adopt hierarchical structures that obtain word-level or utterance-level representations via PLMs and then devise other elaborate modules for knowledge complement. Some of them explicitly combine structured commonsense knowledge to the model and form knowledge-enriched representations (Zhong et al., 2019; Zhang et al., 2020). For the knowledge that is abstractive or unstructured, some other models adopt multi-task learning to compensate for missing knowledge dimensions implicitly (Wang et al.,

* Equal contribution
† Email corresponding

2020; Li et al., 2020).

However, the above works do not consider internal interactions between utterance and knowledge representations when incorporating commonsense knowledge but simply concatenate them, which may negatively affect model performance as proved in the follow-up experiment. Besides, the auxiliary tasks of most multi-task learning methods are emotion-indirect, such as topic inference (Wang et al., 2020) and utterance-speaker verification (Li et al., 2020), which do not involve additional affective information directly. Ilievski et al. (2021) also show that knowledge overlap between different knowledge sources is little. Intuitively for the ERC task, the complement of different dimensions of knowledge helps the reasoning process. In Figure 1, We illustrate a conversation where both commonsense knowledge and sentiment lexicon aid emotion detection. For example, considering the keyword "divorce" in the first utterance, with "an_affair" as a possible cause, "stop_being_married" as an action, and "divide_a_family" as a result, commonsense knowledge enables the model to build a semantics-enhanced chain on "divorce". The sentiment lexicon assigns extremely negative sentiment polarity intensity "-0.83" for "divorce", which directly instructs the model on determining negative emotions. Obviously, in the process of judging this utterance as "Frustrated", the two sources of knowledge have played their respective roles.

To cope with the above challenges, we propose a Knowledge-Interactive Network with sentiment polarity intensity-aware multi-task learning (KI-Net). We first adopt a context- and dependency-aware encoder for context modeling. To further enhance the word-level representations, we leverage a large-scale commonsense knowledge graph and a sentiment lexicon. Inspired by Yang et al. (2019a), knowledge representations are incorporated into word-level representations using a self-matching mechanism, allowing a full internal interaction. We also introduce a phrase-level Sentiment Polarity Intensity Prediction (SPIP) as the auxiliary task, which is expected to provide more direct instructions on emotion recognition of the target utterance.

In summary, this paper makes the following contributions:

- We try to make up for some of the missing knowledge dimensions in the PLM by applying multi-source knowledge. The subsequent ablation study shows that the introduced knowledge does have a positive impact on the performance of the model.
- For the first time on the ERC task, we discuss the necessity of explicit interactions between utterance and knowledge, guiding future work of knowledge integration.
- We adopt a new auxiliary task for ERC, namely phrase-level sentiment polarity intensity prediction. Experiments show that the SPIP task provides promising improvement for the ERC task.

2 Related Work

Emotion recognition in conversation has gained attention from the NLP community only in the past few years (Yeh et al., 2019; Majumder et al., 2019; Zhou et al., 2018) since the growing availability of public conversational data (Busso et al., 2008; Poria et al., 2019a; Li et al., 2017).

ERC task naturally requires modeling interaction between conversation participants. Considering this requirement, many works adopt RNNs to model contextual utterances in a temporal sequence, such as CMN (Hazarika et al., 2018b) and ICON (Hazarika et al., 2018a). Based on them, Majumder et al. (2019) propose a attentive RNN-based model DialogueRNN to model party states and emotional dynamics. Transformer (Vaswani et al., 2017) has also been devised to model input sequences in many recent works (Zhong et al., 2019; Zhang et al., 2020), which lead to better results. Besides, modules such as memory networks (Wenxiang Jiao and King, 2020; Xing et al., 2020) and graph-based networks (Ghosal et al., 2019; Ishiwatari et al., 2020) are also introduced for representation learning to better model contextual information and utterance dependencies.

Limited by the scale of current available high-quality datasets, some works manage to incorporate task-related knowledge to boost model performance. Hazarika et al. (2021); Chapuis et al. (2020) propose elaborate pre-training tasks to improve generalization of models. Zhong et al. (2019); Zhang et al. (2020) explicitly extract commonsense knowledge from large-scale knowledge graphs and concatenate them to word embeddings, forming knowledge-enriched representations. In addition, some works implicitly introduce knowl-

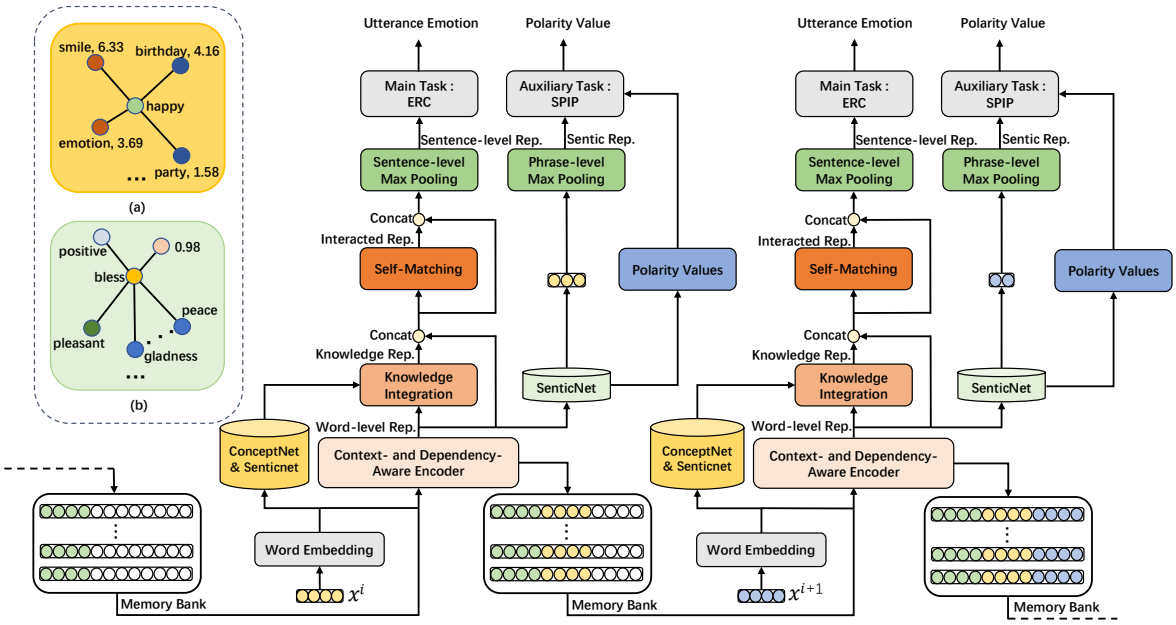

Figure 2: Overall architecture of our model. Rep. denotes representation. (a) is a sub-graph extracted from the ConceptNet with the keyword "happy" while (b) is an example provided by the SenticNet with the keyword "bless".

edge through multi-task learning, such as label imbalance confusion (Zhang et al., 2020), dialogue topic information (Wang et al., 2020) and utterance-speaker relations (Li et al., 2020).

3 Our Proposed KI-Net Model

3.1 Task Definition and Model Overview

We define ERC task as follows: Given $\{\{X_j^i\}, Y^i\}$, where $i = 1, ...N, j = 1, ...N^i$, as a collection of {utterance, emotion label} pairs in one conversation. Conversation $\mathbf{X}$ consists of N utterances and each utterance $\mathbf{X}^i$ consists of N^i tokens, namely $\mathbf{X}^i = \{X_1^i, X_2^i, ..., X_{N_i}^i\}$. Each $\mathbf{X}^i$ is uttered by $p(\mathbf{X}^i) \in \mathbf{P}$, where $\mathbf{P}$ is the set of conversation participants. The discrete value $Y^i \in \mathbf{S}$ is used to denote the emotion label, where $\mathbf{S}$ denotes the set of pre-defined emotion labels, and $|\mathbf{S}| = h_c$. The objective of the ERC task is to predict the emotion label Y^i of the target utterance $\mathbf{X}^i$ given its previous context and the mappings between $\mathbf{X}$ and $\mathbf{P}$. Our proposed KI-Net is illustrated in Figure 2.

To aid lacking knowledge dimensions of PLM, we design a hierarchical model, whose key idea is to enhance PLM with rich-interacted and strongly-correlated knowledge. Based on this idea, KI-Net is built, as depicted in Figure 2, with four major components. We first use a XLNet-based encoder that computes context- and dependency-aware representations for utterances (Sec. 3.2). Then a knowledge introduction module is devised to retrieve commonsense knowledge and form graph attention-based representation (Sec. 3.3). A self-matching module is employed for utterance-knowledge interaction based on self-attention mechanisms (Sec. 3.4). We also propose a SPIP task, which introduces strongly-correlated knowledge to the model, and a multi-task learning setting to combine ERC and SPIP task (Sec. 3.5).

3.2 Context- and Dependency-Aware Encoder

Both historical conversational information and dependency modeling are crucial for the ERC task. Therefore, based on XLNet (Yang et al., 2019b), we use a Context- and Dependency-Aware (CDA) encoder to exploit both of the above elements by improving the original self-attention mechanism.

For the time step i, the target utterance $\mathbf{X}^i$ is prepended with the "$[CLS]$" token: $\mathbf{x}^i = \{[CLS], X_1^i, X_2^i, ..., X_{N_i}^i\}$. Then $\mathbf{x}^i$ is passed through the embedding layer:

$$\mathbf{h}_0^i = \text{embedding}(x^i) \quad (1)$$

where $\mathbf{h}_0^i \in \mathbb{R}^{N^i \times D_h}$, and D_h denotes input dimension of XLNet-base. $\mathbf{h}_0^i$ is regarded as input states of the CDA encoder's first layer. Also, $\mathbf{h}_0^i$ is used in concept embedding layer of knowledge introduction, which we will discuss in Sec. 3.3.

Besides the ordinary global self-attention, our model devises a local self-attention which uses a limited conversational window size to focus on the neighboring part of the target utterance, a speaker self-attention which retains historical context belonging to the target speaker and listener self-attention which focuses on historical context uttered by the other participants. These four types of self-attention results are combined to form the output of every layer in the CDA encoder. Following DialogXL (Shen et al., 2020), the context memory $\mathbf{m}$ is combined with hidden states using a utterance recurrence mechanism. Given the input $\mathbf{h}_0^i$, the CDA encoder adopts L layers of Transformer to get word-level representation. For convenience, we denote this process as:

$$\mathbf{h}_L^i = \text{encoder}(\mathbf{h}_0^i, \mathbf{m}^{i-1}) \tag{2}$$

where $\mathbf{h}_L^i \in \mathbb{R}^{N^i \times D_h}$, and $\mathbf{m}^{i-1} \in \mathbb{R}^{L \times D_m \times D_h}$, D_m is a pre-defined max memory length. encoder denotes the encoding process.

3.3 Knowledge Introduction

This section introduces the knowledge introduction process where ConceptNet (Speer et al., 2017) is leveraged as the commonsense knowledge base. ConceptNet is a large multi-lingual semantic graph, where each node denotes a phrase-level concept and each edge denotes a relation. Each quadruple <concept1, relation, concept2, weight> in ConceptNet denotes an assertion, where the weight is a confidence score assigned to the assertion.

We first introduce the knowledge retrieval process. For a token t, we extract a graph $\mathbf{g}_t$, which consists of t's immediate neighbors from ConceptNet. For each $\mathbf{g}_t$, we discard concepts that are stopwords or not in the word vocabulary $\mathbf{V}$ of the encoder mentioned in last section, and remove assertions with confidence scores less than 1.0 for denoising. $\mathbf{g}_t = \{(c_1, w_1), (c_2, w_2), ..., (c_k, w_k)\}$, where c_i denotes the i^{th} connected concept of t, w_i denotes its corresponding confidence score. An example of $\mathbf{g}_t$ is illustrated in Figure 2 (a).

We then adopt a graph attention mechanism to form knowledge representations. For each non-stop token $X_j^i \in \mathbf{x}^i$, we have a graph $\mathbf{g}_j^i$. For X_j^i and $c_p \in \mathbf{g}_j^i$, we obtain their embedding $\mathbf{h}_0^{ij}$ and $\mathbf{h}_0^{c_p}$ via embedding layer mentioned in Equation.1. Then knowledge representation $\mathbf{k}_j^i$ are computed as follows:

$$t_p = \mathbf{h}_0^{ij} \cdot \mathbf{h}_0^{c_p} \tag{3}$$

$$\alpha_p = \frac{\exp(t_p w_p)}{\sum_{p=1}^{N_c^{ij}} \exp(t_p w_p)} \tag{4}$$

$$\mathbf{k}_j^i = \sum_{p=1}^{N_c^{ij}} \alpha_p \mathbf{h}_0^{c_p} \tag{5}$$

where $\mathbf{h}_0^{ij}, \mathbf{h}_0^{c_p}, \mathbf{k}_j^i \in \mathbb{R}^{D_h}$, $\cdot$ denotes dot product operation, and N_c^{ij} denotes the number of concepts in $\mathbf{g}_j^i$. If $N_c^{ij} = 0$, we set $\mathbf{k}_j^i$ to the average of all node vectors.

We adopt SenticNet (Cambria et al., 2020) as another knowledge source. For each phrase s^i in SenticNet, there is a quintuple <polarity_value, polarity_intense, moodtags, sentics, semantics>, where the polarity_value belongs to positive or negative. Polarity_intense is a floating number between -1 and +1, denoting positivity of s^i. For phrase s^i, its mood tags $\hat{\mathbf{m}}^i \subset \mathbf{M}$, where $\mathbf{M}$ is the set of pre-defined emotion description words. Sentics is a quadruple and semantics $\hat{\mathbf{e}}^i$ defines a set of semantics-related concepts of s^i. An example of these tuples is illustrated in Figure 2 (b).

We add mood tags and semantics to the commonsense knowledge base retrieved in Sec. 3.3. Specifically, for $s^i \in \mathbf{V}$, we construct a mood tag set with a weight value $\hat{\mathbf{m}}^i = \{(m_1^i, w_0), (m_2^i, w_0), ..., (m_{N_m^i}^i, w_0)\}$, where $w_0 = 2.0$, N_m^i is the number of mood tags in $\hat{\mathbf{m}}^i$. Similarly, we have a semantics set with weight $\hat{\mathbf{e}}^i = \{(e_1^i, \hat{w}_0), (e_2^i, \hat{w}_0), ..., (e_{N_e^i}^i, \hat{w}_0)\}$, where $\hat{w}_0 = 1.0$, N_e^i is the number of semantics in $\hat{\mathbf{e}}^i$. With $\hat{\mathbf{m}}^i$ and $\hat{\mathbf{e}}^i$, we construct enhanced knowledge graph as follows: $\hat{\mathbf{g}}_{s^i} = \mathbf{g}_{s^i} \cup \hat{\mathbf{m}}^i \cup \hat{\mathbf{e}}^i$, where $\cup$ denotes union operation of sets.

With enhanced knowledge graph $\hat{\mathbf{g}}$, we make minor changes to Equation. 3. For X_j^i and $\hat{c}_p \in \hat{\mathbf{g}}_j^i$, we obtain their token embeddings $\mathbf{h}_0^{ij}$ and $\mathbf{h}_0^{\hat{c}_p}$ via embedding layer mentioned in Equation.1. We modify Equation. 3 as follows:

$$t_p = \mathbf{h}_0^{ij} \cdot \mathbf{h}_0^{\hat{c}_p} \tag{6}$$

where $\mathbf{h}_0^{\hat{c}_p} \in \mathbb{R}^{D_h}$. Then t_p is used for computation of Equation. 4, with the rest unchanged.

3.4 Self-Matching

To employ internal utterance-knowledge interaction in the model, we propose a self-matching module based on self-attention. For each token X_j^i, we

obtain $\mathbf{u}_j^i$ as follows:

$$\mathbf{u}_j^i = [\mathbf{h}_L^{ij}; \mathbf{k}_j^i] \tag{7}$$

where [;] denotes concatenation operation, $\mathbf{h}_L^{ij} \in \mathbb{R}^{D_h}$ and $\mathbf{u}_j^i \in \mathbb{R}^{2D_h}$. For two tokens X_j^i and X_m^i within one utterance, we compute their similarity via a trilinear function (Seo et al., 2017):

$$\hat{r}_m^j = \mathbf{W}^{\mathrm{T}}[\mathbf{u}_j^i; \mathbf{u}_m^i; \mathbf{u}_j^i \odot \mathbf{u}_m^i] \tag{8}$$

where $\mathbf{W} \in \mathbb{R}^{6D_h}$ is the model parameter, and $\odot$ denotes element-wise multiplication. We obtain the similarity matrix $\hat{\mathbf{R}}$ accordingly with $\hat{r}_m^j$ being the jm^{th} entry. Then we obtain the self-attention matrix $\mathbf{Q}$ as follows:

$$q_m^j = \frac{\exp(\hat{r}_m^j)}{\sum_m^{N_i} \exp(\hat{r}_m^j)} \tag{9}$$

where q_m^j is the jm^{th} entry of $\mathbf{Q}$.

Intuitively, indirect interaction allows the model to learn deeper semantic relations within the knowledge-enriched representations. To further achieve indirect interaction, we conduct a self-multiplication of $\mathbf{Q}$:

$$\hat{\mathbf{Q}} = \mathbf{Q}\mathbf{Q}^{\top} \tag{10}$$

With indirect interaction, all token pairs can interact through every other token within the utterance.With $\mathbf{Q}$ and $\hat{\mathbf{Q}}$, we compute two attended vectors for each token X_j^i:

$$\mathbf{v}_j^i = \sum_m^{N_i} q_m^j \mathbf{u}_m^i \tag{11}$$

$$\hat{\mathbf{v}}_j^i = \sum_m^{N_i} \hat{q}_m^j \mathbf{u}_m^i \tag{12}$$

where $\mathbf{v}_j^i, \hat{\mathbf{v}}_j^i \in \mathbb{R}^{2D_h}$, $\hat{q}_m^j$ is the jm^{th} entry of $\hat{\mathbf{Q}}$. We concatenate the two attended vectors with different means to allow rich interactions:

$$\mathbf{c}_j^i = [\mathbf{u}_j^i; \mathbf{v}_j^i; \mathbf{u}_j^i - \mathbf{v}_j^i; \mathbf{u}_j^i \odot \mathbf{v}_j^i; \hat{\mathbf{v}}_j^i; \mathbf{u}_j^i - \hat{\mathbf{v}}_j^i] \tag{13}$$

where $\mathbf{c}_j^i \in \mathbb{R}^{12D_h}$, and $\mathbf{c}_j^i$ denotes the j^{th} row of self-matching output matrix $\mathbf{C}$. $\mathbf{C}$ is derived by semantics and knowledge interactions between utterance tokens, which allows knowledge to be introduced purposefully instead of acting as noise.

3.5 Sentiment Polarity Intensity Prediction

In this section, we propose a phrase-level Sentiment Polarity Intensity Prediction (SPIP) task. SPIP is used as an auxiliary task to incorporate sentiment polarity knowledge to the model. Specifically, the model predicts sentiment intensive values for all SenticNet phrases within the utterance. For $\mathbf{x}^i$, we retrieve a set $\mathbf{P}^i = \{\mathbf{p}_k^i | \mathbf{p}_k^i \in$ n-grams from $\mathbf{x}^i\}, n = 1, 2, ..., N_g$, where N_g is a hyper-parameter. For $\mathbf{p}_k^i \in$ SenticNet $\cap\, V$, where $\mathbf{p}_k^i$ denotes k^{th} phrase of $\mathbf{P}^i$, we record their starting and ending positions $< P_0^{ik}, P_1^{ik} >$ in the utterance, and the corresponding intensive value O_k^i. Therefore, for each utterance x^i we have $\{< P_0^{ik}, P_1^{ik} >, O_k^i\}$, k=1,...,$\hat{N}_i$, where $\hat{N}_i$ denotes the number of SenticNet phrases within $\mathbf{x}^i$.

For each utterance $\mathbf{x}^i$, we obtain its word-level representation $\mathbf{h}_L^i$ via Equation.2. For SenticNet phrase $\mathbf{p}_k^i$, we obtain its representation $\mathbf{r}_k^i$ using phrase-level max pooling:

$$\hat{\mathbf{h}}_k^i = \mathbf{h}_L^i[P_0^{ik} : P_1^{ik}] \tag{14}$$

$$\mathbf{r}_k^i = \mathrm{maxpooling}(\hat{\mathbf{h}}_k^i \mathbf{W}_0 + \mathbf{b}_0) \tag{15}$$

where $\hat{\mathbf{h}}_k^i \in \mathbb{R}^{|P_1^{ik} - P_0^{ik}| \times D_h}$, $\mathbf{r}_k^i \in \mathbb{R}^{D_h}$, $\mathbf{W}_0 \in \mathbb{R}^{D_h \times h}$ and $\mathbf{b}_0 \in \mathbb{R}^h$ are model parameters, h denotes a pre-defined hidden dimension, [:] denotes matrix slice operation, and maxpooling denotes the max pooling operation. We compute the final prediction score:

$$\hat{O}_k^i = \tanh(\mathbf{r}_k^i \mathbf{W}_1 + \mathbf{b}_1) \tag{16}$$

where $\mathbf{W}_1 \in \mathbb{R}^{h \times 1}$ and $\mathbf{b}_1 \in \mathbb{R}^1$ are model parameters. As training objective, we compute standard MSE loss for SPIP task:

$$loss_a = \frac{1}{N * \hat{N}_i} \sum_{i=1}^{N} \sum_{k=1}^{\hat{N}_i} (\hat{O}_k^i - O_k^i)^2 \tag{17}$$

For utterance $\mathbf{x}^i$, we have obtained its word-level knowledge-enriched representations $\mathbf{c}^i$ from self-matching layer (Sec. 3.4), where $\mathbf{c}^i$ is the i^{th} entry of $\mathbf{C}$. We compute its utterance-level representation through max pooling:

$$\hat{\mathbf{c}}^i = \mathrm{maxpooling}(\mathbf{c}^i \mathbf{W}_2 + \mathbf{b}_2) \tag{18}$$

where $\mathbf{c}^i \in \mathbb{R}^{N^i \times 12D_h}$, $\mathbf{W}_2 \in \mathbb{R}^{12D_h \times h}$ and $\mathbf{b}_2 \in \mathbb{R}^h$ are model parameters. We compute final classification probabilities as follows:

$$\hat{\mathbf{Y}}^i = \mathrm{softmax}(\hat{\mathbf{c}}^i \mathbf{W}_3 + \mathbf{b}_3) \tag{19}$$

Dataset	Conv.(Train/Val/Test)	Utter.(Train/Val/Test)	Utter./Conv.
IEMOCAP	100/20/31	4,810/1,000/1,623	49.2
MELD	1,038/114/280	9,989/1,109/2,610	9.6
DailyDialog	11,118/1,000/1,000	87,170/8,069/7,740	7.9

Table 1: The statistics of the datasets.

where $\hat{\mathbf{c}}^i \in \mathbb{R}^h$, $\mathbf{W}_3 \in \mathbb{R}^{h \times h_c}$ and $\mathbf{b}_3 \in \mathbb{R}^{h_c}$ are model parameters. $\mathrm{softmax}$ denotes the softmax operation.

We compute the loss of ERC task using standard cross-entropy loss:

$$loss_m = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{h_c}\Big(Y^i log\hat{Y}_k^i + (1-Y^i)log(1-\hat{Y}_k^i)\Big) \quad (20)$$

With both $loss_m$ for the main task ERC and $loss_a$ for auxiliary task SPIP, we compute the total loss of the task as follows:

$$loss = \frac{loss_m + \epsilon loss_a}{1+\epsilon} \quad (21)$$

where $\epsilon \in [0, 1]$ is the pre-defined weight coefficient of $loss_a$.

4 Experimental Setting

In this section we present experimental settings such as datasets, baselines, implementation details and evaluation metrics.

4.1 Datasets

We evaluate our model on the following three ERC datasets. The statistics are shown in Table 1.

IEMOCAP (Busso et al., 2008): A multi-modal conversation dataset, with emotion labels neutral, happiness, sadness, anger, frustrated and excited. Each conversation includes two parties.

MELD (Poria et al., 2019a): A multi-modal dataset enriched from $EmotionLines$ dataset, collected from the scripts of TV show $Friends$. The labels are neutral, happiness, surprise, sadness, anger, disgust and fear.

DailyDialog (Li et al., 2017): From human-written daily conversations with no speaker information. The labels are similar to MELD.

4.2 Baselines and State of the Art

We compare our model with the following baselines:

BERT_BASE (Devlin et al., 2019): Initialized with pre-trained parameters of BERT-BASE, the model is fine-tuned for ERC task.

DialogueRNN (Majumder et al., 2019): DialogRNN uses three GRUs to model speaker states, global contexts and emotion dynamics. The model is expected to model inter-speaker relations on multi-party conversations.

DialogueGCN (Ghosal et al., 2019): The model utilizes a graph-based structure to model utterance relations within a conversation.

KET (Zhong et al., 2019): The model uses a graph attention mechanism to combine commonsense knowledge into utterance representations.

AGHMN (Wenxiang Jiao and King, 2020): The model uses a hierarchical memory network to model and store context representations.

HiTrans (Li et al., 2020): Based on hierarchical Transformer, the model uses multi-task learning to be speaker-sensitive.

IEIN (Lu et al., 2020): IEIN uses predicted emotion labels instead of gold labels and designs a loss to constrain the prediction of each iteration.

RGAT (Ishiwatari et al., 2020): Based on graph structure, the model augments relation modelling of conversations, and adds relational position encodings to combine sequential information.

COSMIC (Ghosal et al., 2020): COSMIC incorporates different elements of commonsense and leverages them to learn interlocutors' interactions.

DialogXL (Shen et al., 2020): The model uses a dialog-aware self-attention to introduce the awareness of inter- and intra-speaker dependency.

4.3 Other Experimental Settings

We conducted all experiments using Xeon(R) Silver 4110 CPU with 768GB of memory and GeForce GTX 1080Ti GPU with 11GB of memory. We tokenize and pre-process the above three datasets and use the XLNet tokenizer provided by Hugging Face [1] to correspond with the vocabulary of the pre-trained XLNet. For hyper-parameter setting, D_h=768, h=300, L=12, N_g=4, h_c and D_m depends on the dataset. We set the initial learning rates of 1e-5 on IEMOCAP, 1e-6 on MELD and DailyDialog. We employ AdamW optimizer (Loshchilov and Hutter, 2017) the scheduled learning rate with a batch size of 6 on on IEMOCAP and 4 on MELD and DailyDialog during training. WE set 0.3 as the dropout rate on DailyDialog and 0 on the rest dataset. All the results are obtained using the text modality only. The evaluation metrics are chosen as micro-F1 for DailyDialog and

[1]The website of Hugging Face: https://huggingface.co/

Model	Happy	Sad	Neutral	Angry	Excited	Frustrated	Avg.
BERT_BASE (Devlin et al., 2019)	37.09	59.53	51.73	54.33	54.26	55.83	53.31
DialogueRNN (Majumder et al., 2019)	33.18	**78.8**	59.21	**65.28**	**71.86**	58.91	62.75
DialogueGCN (Ghosal et al., 2019)	42.75	**84.54**	63.54	64.19	63.08	**66.99**	64.18
KET (Zhong et al., 2019)	–	–	–	–	–	–	59.56
AGHMN (Wenxiang Jiao and King, 2020)	**52.10**	73.30	58.40	61.90	69.70	62.30	63.50
HiTrans (Li et al., 2020)	–	–	–	–	–	–	64.50
IEIN (Lu et al., 2020)	**53.17**	77.19	61.31	61.45	69.23	60.92	64.37
RGAT (Ishiwatari et al., 2020)	51.62	77.32	**65.42**	63.01	67.95	61.23	65.22
COSMIC (Ghosal et al., 2020)	–	–	–	–	–	–	65.28
DialogXL (Shen et al., 2020)	–	–	–	–	–	–	65.94
KI-Net + BERT	39.10	65.24	57.35	57.81	60.17	57.61	59.93
KI-Net + XLNet	47.63	72.47	63.88	64.0	69.40	62.02	64.72
KI-Net (Ours)	49.45	73.38	**65.63***	**65.13**	**71.15**	**68.38***	**66.98***

Table 2: Performance comparison of ours, baselines, and state-of-the-art method for each emotion and their averages on IEMOCAP. We highlight top-2 values on each emotion in bold. "–" means the original paper do not give the corresponding result. The numbers with * indicate that the improvement of our model over all baselines is statistically significant with $p < 0.05$ under t-test.

Model	MELD	DailyDialog
BERT_BASE	56.21	53.12
DialogueRNN	57.03	50.65
DialogueGCN	58.10	–
KET	58.18	53.37
AGHMN	58.10	–
HiTrans	61.94	–
IEIN	60.72	54.71
RGAT	60.91	54.31
COSMIC	**65.21**	**58.48**
DialogXL	62.41	54.93
KI-Net + BERT	60.60	54.33
KI-Net + XLNet	62.12	55.07
KI-Net (Ours)	**63.24**	**57.30**

Table 3: Performance comparisons on MELD and DailyDialog. We highlight top-2 values in bold.

weighted-F1 for the other datasets. The results reported in our experiments are all based on average of 5 random runs on the test set.

5 Results and Analysis

5.1 Overall Results

Overall results of our model and the baselines are listed in Table 2 and Table 3. According to the results on IEMOCAP. DialogXL, COSMIC and RGAT outperform other models with a performance of over 65%. All these three models consider modeling dependencies within conversations, indicating that elaborate context modeling modules are essential for the ERC task again. We also notice that models such as KET improve transformer-based PLM since they explicitly introduce commonsense knowledge. Besides, HiTrans devises an auxiliary task to combine task-related information, which also shows some improvement. Our KI-Net model refreshed the current best performance on IEMOCAP, bringing a 1.04% performance improvement. We attribute this result to our model considering all the three factors mentioned above.

Similar results are also reflected on MELD and DailyDialog. KI-Net achieves 63.24% on MELD, which is around 5% better than KET. Considering the structure of KET, we believe that this improvement mainly comes from the introduction of self-matching modules. KI-Net achieves 57.30% on DailyDialog, which is around 2.5% better than DialogXL. This may because external knowledge complements the lacking knowledge dimensions of PLMs. KI-Net is weaker than COSMIC on these two datasets while still ranks in the top-2 positions. Unlike our model, COSMIC leverages a different set of PLM and knowledge source. We speculate that the performance on short conversations (less than ten turns) will be more dependent on the selection of knowledge sources.

5.2 Emotion-Specific Results

We present emotion-specific testing results on the IEMOCAP dataset in Table 2. KI-Net remains top-2 for most emotions and shows a balanced performance. Specifically, on emotions Neutral and Frustrated, our model achieves the best results at 65.63% and 68.38%. We believe the in-

Figure 3: Two cases from the IEMOCAP dataset. Golden, CDA. and w/o SPIP denote the ground-truth label, the prediction of CDA encoder and KI-Net without SPIP. The boxes linked to w/o SPIP and KI-Net denote the attention weights of the top-8 attended concepts of the key token and the polarity_intense prediction results respectively.

Method	IEMOCAP	MELD	DailyDialog
sentiment intensive value	66.98	**63.24**	**57.3**
sentiment polarity	**67.20**	62.93	57.24
mood tags	66.63	62.87	57.0

Table 4: Results of Different Elements for SPIP.

teraction between the knowledge and the utterance provides reasonable instructions on the final judgment, which benefits fine-grained emotions' detection such as Frustrated and Angry.

5.3 Effect of Element Selection for SPIP

As mentioned above, for each phrase s^i in SenticNet, there is a tuple with some sentiment-related elements. In addition to the sentiment intensive value, we also explore some of the other elements to provide supervision information for our auxiliary tasks. The results are shown in Table 4. We tried different combinations, such as the sentiment polarity and mood tags, but the effect is weaker than sentiment intensity. We think this is because sentiment intensity already includes sentiment polarity, and SPIP is a phrase-level auxiliary task, but the main task needs to be judged by context, which will shift the fine-grained emotions corresponding to mood tags, so sentiment intensity is a more appropriate choice.

5.4 Case Study

We provide two cases obtained from the actual testing process of the IEMOCAP dataset to verify the influence brought by the introduced knowledge and the SPIP task. As illustrated in Figure 3, in case 1, with commonsense concepts such as "miss", "husband" and "wedded" etc, the model gains more profound insight into the semantics of "married". Meanwhile, the SPIP classifier gives relative strong positivity for the phrase "getting_married", which establishes the emotional direction of the target utterance with another keyword "not". Obviously, these two ways of knowledge introduction play different roles in the reasoning process. The result of the CDA encoder further shows that context plays little role in this case.

In case 2, we can see the model does not get direct emotion-related information via commonsense knowledge concepts such as "earned". Hence, in this case, the knowledge introduction module plays a relatively little role and makes the same prediction as to the CDA encoder. However, with the negative intensity value that the SPIP classifier gives to the token "cheap", the model manages to label the utterance "Angry", which is also a negative emotion but obviously more intensive than "Frustrated".

5.5 Error Analysis

We present the confusion matrix of our test results on the IEMOCAP dataset in Figure 4. We notice that many of the misclassifications are between neutral and non-neutral emotions which can be improved by adding clues from other modalities. Despite the strong performance of our model, it still shows that distinguish similar emotions (e.g., excited and happy) remains a great challenge. A possible reason is that the sentiment lexicon assigns close polarity intense values to words with similar emotional expressions.

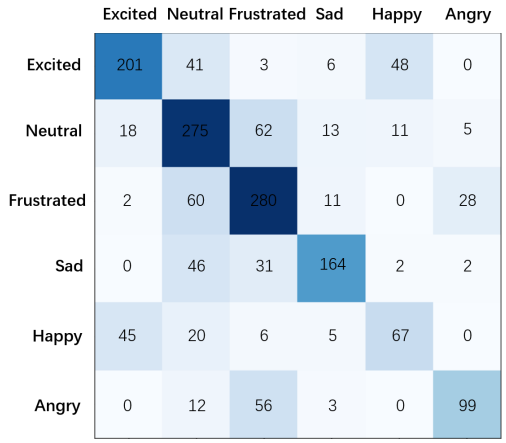

Figure 4: Confusion matrix on IEMOCAP.

5.6 Ablation Study

We perform an ablation study for our designed modules. "-Self-Matching" denotes that the utterance and knowledge representation are directly concatenated. "-Knowledge Integration" means both knowledge introduction and self-matching are discarded. As shown in Table 5, the performance drops with any of the components removed. Especially after deleting self-matching, the performance may even lower than the CDA encoder. This result proves that self-matching is crucial for integrating knowledge, without which knowledge may even bring the noise to emotional reasoning.

Performance drops more when the SPIP is removed on the IEMOCAP dataset while knowledge integration plays a relatively more important role on the other two datasets. This may because there is only an average of 1.9 Sentic phrases per utterance with a 65% probability on the MELD dataset. To some extent, this once again confirms our previous conjecture that short conversations are more critical of knowledge sources than long conversations.

Method	IEMOCAP	MELD	DailyDialog
KI-Net	**66.98**	**63.24**	**57.3**
-Knowledge Integration	66.58 (↓ 0.40)	62.72 (↓ 0.52)	56.52 (↓ 0.78)
-Self-Matching	64.89 (↓ 2.09)	62.38 (↓ 0.86)	55.35 (↓ 1.95)
-SPIP	66.39 (↓ 0.59)	62.89 (↓ 0.35)	57.07 (↓ 0.23)
CDA encoder	65.88 (↓ 1.10)	62.42 (↓ 0.82)	54.82 (↓ 2.48)

Table 5: Results of ablation study.

6 Conclusion

This paper proposes a KI-Net for emotion recognition in conversations. Our model outperforms state-of-the-art models on IEMOCAP. Extensive experiments prove the necessity of interaction between knowledge and utterance, and the new auxiliary task SPIP will further improve performance. Utterance-level interaction and confusion of similar emotions are the focus of our following research. Which dimensions of knowledge ERC relies more on is also worthy of in-depth discussion.

Acknowledgements

We thank Shaobo Li for his insightful discussions. We also thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Key R&D Program of China via grant 2020YFB1406902.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 105–114, New York, NY, USA. Association for Computing Machinery.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

A. Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, R. Srikanth, Michel Galley, and P. Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Comput. Hum. Behav.*, 93:309–317.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2021. Conversational transfer learning for emotion recognition. *Information Fusion*, 65:1–12.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Filip Ilievski, A. Oltramari, Kaixin Ma, B. Zhang, D. McGuinness, and Pedro A. Szekely. 2021. Dimensions of commonsense knowledge. *ArXiv*, abs/2101.04640.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370, Online. Association for Computational Linguistics.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Runnan Li, Zhiyong Wu, Jia Jia, Yaohua Bu, Sheng Zhao, and Helen Meng. 2019. Towards discriminative representation learning for speech emotion recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5060–5066. International Joint Conferences on Artificial Intelligence Organization.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

I. Loshchilov and F. Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4078–4088.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings*

of the AAAI Conference on Artificial Intelligence, 33(01):6818–6825.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and E. Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. *ArXiv*, abs/1611.01603.

Weizhou Shen, J. Chen, Xiaojun Quan, and Zhixian Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *ArXiv*, abs/2012.08695.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. *ArXiv*, abs/2004.05483.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9177–9184.

Michael R. Lyu Wenxiang Jiao and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8002–8009.

S. Xing, S. Mai, and H. Hu. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, pages 1–1.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019a. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

S. Yeh, Y. Lin, and C. Lee. 2019. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. 2020. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4429–4440, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, T. Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.