SKIER: A Symbolic Knowledge Integrated Model for Conversational Emotion Recognition

Wei Li, Luyao Zhu, Rui Mao, Erik Cambria

School of Computer Science and Engineering, Nanyang Technological University, Singapore.
{wei008, luyao001}@e.ntu.edu.sg, {rui.mao, cambria}@ntu.edu.sg

Abstract

Emotion recognition in conversation (ERC) has received increasing attention from the research community. However, the ERC task is challenging, largely due to the complex and unstructured properties of multi-party conversations. Besides, the majority of daily dialogues take place in a specific context or circumstance, which requires rich external knowledge to understand the background of a certain dialogue. In this paper, we address these challenges by explicitly modeling the discourse relations between utterances and incorporating symbolic knowledge into multi-party conversations. We first introduce a dialogue parsing algorithm into ERC and further improve the algorithm through a transfer learning method. Moreover, we leverage different symbolic knowledge graph relations to learn knowledge-enhanced features for the ERC task. Extensive experiments on three benchmarks demonstrate that both dialogue structure graphs and symbolic knowledge are beneficial to the model performance on the task. Additionally, experimental results indicate that the proposed model surpasses baseline models on several indices.

Introduction

Emotion recognition in conversation (ERC) is a task that is beneficial to a wide range of natural language processing (NLP) research domains, such as dialogue systems (Ma et al. 2020) and sentiment analysis (Zhang et al. 2021). ERC is featured by the fact that the emotion classification task depends on both current and historical utterances from different speakers. Thus, unlike phrase-level (Ge, Mao, and Cambria 2022), aspect-level (Liang et al. 2022; Mao and Li 2021), sentence-level (Chen et al. 2017) and document-level (Zhao, Rao, and Feng 2017) affective computing tasks modeling dependency relationships within a context given by a single presenter, utterance-level ERC requires modeling the various dependencies across multiple speakers.

Previous ERC studies have formulated two main trends, e.g., sequence-based methods and graph-based methods (Li et al. 2020a). The former trend (Song et al. 2022) encoded concatenated historical and current utterances, and predicted an emotion class for the current utterance, based on contextualized encoders, e.g., LSTM (Hochreiter and Schmidhuber 1997), GRU (Cho et al. 2014) and pre-trained language models (Devlin et al. 2019; Liu et al. 2019).

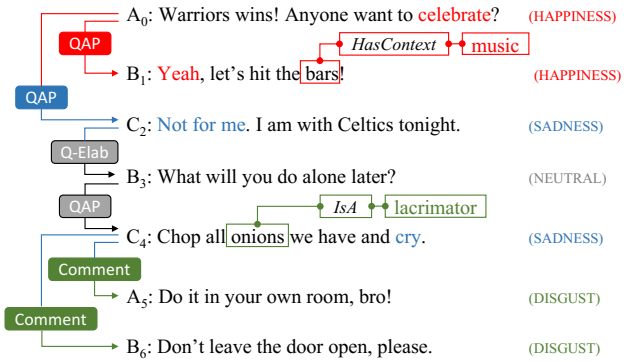

Figure 1: Discourse with symbolic dependency representations. A, B, and C are three different speakers. QAP is a question-answer pair. Q-Elab is a question-elaboration.

The later trend (Ghosal et al. 2020b) used graph convolutional networks (GCNs) (Kipf and Welling 2017) to model historical and current utterance context, and utterance-speaker relationships. The dependencies were represented as nodes and edges in a graph. The cross utterance dependency modeling of the above-mentioned technical trends was normally achieved by contextualizers and attention mechanisms in vector space. Despite the fact that contextualizer-based methods have significantly improved ERC by incorporating more contextual information represented in sequential or graphic forms, using attention mechanisms and undirected edges in a graph cannot model the diversity of dialogue dependencies (Xu et al. 2019).

We argue that explicitly learning different types of dependencies can deliver extra accuracy gains in learning ERC and dependency explainability in ERC results. Given the hypotheses that (1) symbolic dependency representations can represent different types of dependency relationships between utterances (Shi and Huang 2019); (2) commonsense knowledge can help to infer the emotion class of an utterance from its context (Zhong, Wang, and Miao 2019), the motivation of this work is to effectively use and fuse these different kinds of symbolic knowledge in an ERC model. For example, as seen in Fig. 1, B_1 and C_2 depend on A_0 in a *QAP* (question-answer pair) relationship, respectively.

Thus, the HAPPINESS emotion of B_1 can inherit from that of A_0 via his agreement response, although there is no emotional word in B_1. Similarly, the SADNESS emotion of C_2 can be inferred from his disagreement to the question of A_0. The agreement (B_1) and disagreement (C_2) utterances to the parent utterance (A_0) in a *QAP* relationship are likely helpful for predicting their emotion classes. The emotional inheritance is not identical in all dependency relationships, e.g., (C_2, B_3) and (B_3, C_4). Thus, it is important to differentiate dependency types between utterances by context. Besides, without commonsense knowledge, e.g., an onion is a type of lacrimator (<*onion*, IsA, *lacrimator*>), and the dependency, e.g., *Comment* in (C_4, A_5) and (C_4, B_6), a classifier can hardly infer the emotions of A_5 and B_6 are DISGUST, because there is no such information, indicating chopping onions are disgusting for people in the context.

In this work, we develop a neurosymbolic model for ERC that leverages the strengths of both deep neural networks and symbolic representations. In particular, our ERC model integrates symbolic dependency knowledge, concept-level commonsense, and sentiment knowledge. The symbolic dependency representations (a discourse graph) are given by a dependency parser proposed by Shi and Huang (2019). To further improve the parser performance in a different conversation domain, we conduct transfer learning (TL) by manually labeling randomly selected seed conversations in ERC datasets and fine-tuning the parser with the seed data. The commonsense knowledge comes from ConceptNet (Speer, Chin, and Havasi 2017) and SenticNet (Cambria et al. 2022). To leverage the multi-level symbolic-based knowledge, we propose a novel graph fusion method. The method integrates concept-level knowledge with a novel attention mechanism and utterance-level knowledge with relational graph convolutional networks (Schlichtkrull et al. 2018).

We employ a RoBERTa (Liu et al. 2019) to enhance contextual and speaker dependency learning. Finally, we use a convolutional self-attention (Dai et al. 2021) to fuse the multi-level symbolic knowledge. We test our model on DailyDialog (Li et al. 2017), Emory (Zahiri and Choi 2018), and MELD (Poria et al. 2019). We focus on ERC from texts, because this is the most fundamental modality in affective computing. We benchmark with state-of-the-art baselines, showing that our method outperforms these baselines by 1.74% on average. We also experimentally demonstrate that both the structured graph-based dependency representations and commonsense knowledge are beneficial to the model performance on the task. The contribution of this work can be summarized as follows: (1) We propose a symbolic knowledge integrated model for the ERC task, named SKIER[1], which effectively leverages symbolic-based dependency knowledge at the utterance level, and commonsense knowledge at the concept level; (2) We introduce a dialogue relation graph-based contextualizer for SKIER to functionally fuse utterance dependencies. Meanwhile, we propose a relation-aware concept representation mechanism to integrate the concepts in different relations; (3) Our method achieves state-of-the-art performance on the ERC task.

[1]https://github.com/senticnet/SKIER

Related Work

There are two technical trends in ERC, namely sequence-based, and graph-based methods (Li et al. 2020a).

Sequence-Based Methods used encoders and attention to learn local and global dependencies (Majumder et al. 2019; Sap et al. 2019; Vaswani et al. 2017; Zhang et al. 2020; Ghosal et al. 2020a; Shen et al. 2021a; Song et al. 2022). Majumder et al. (2019) proposed a GRU-based model, modeling the interactions between speakers, historical context, and historical emotions. Attention was employed to learn the contextual dependency for the speaker states. Li et al. (2020a) proposed a Transformer (Vaswani et al. 2017)-based model. The local utterance representations were given by BERT. A higher level Transformer was employed to learn the global context. Since Transformer is a multi-head attention-based encoder, the dependency of utterances was also modeled by attention. Shen et al. (2021a) fitted utterances into XLNet (Yang et al. 2019) with improved memory efficiency. They also proposed dialog-aware self-attention to learn the intra- and inter-speaker dependencies.

Graph-Based Methods used GCNs to model the relation between utterances and speakers or fuse external knowledge (Zhang et al. 2019; Zhong, Wang, and Miao 2019; Ghosal et al. 2020b). Zhang et al. (2019) introduced a GCN model to leverage both context- and speaker-sensitive dependencies. The utterances and speakers were represented as nodes. The edges represented the dependencies between utterances and the dependencies between utterances and nodes. GCN was used to learn the undirected graph. Zhong, Wang, and Miao (2019) proposed a model that integrates commonsense (ConceptNet) and sentiment (valence, arousal, and dominance, given by NRC VAD (Mohammad 2018)) knowledge. The dependency learning and commonsense fusion were achieved with multiple attention mechanisms, e.g., dynamic context-aware affective graph attention, hierarchical self-attention, and context-response cross-attention. Ghosal et al. (2020b) proposed a DialogueGCN model to learn the intra- and inter-speaker dependencies. The dependency relationship is represented by an edge, connecting past and future utterances within a window to a current utterance.

Although the above contextualizers have significantly improved ERC, they did not explicitly distinguish different dependency types in discourse. For example, the sequence-based methods learn dependencies as the similarity weights between vectors via attention; The graph-based methods represent the dependencies as nodes and edges, and learn the graph via GCNs. We argue that explicitly learning different types of dependencies can deliver extra accuracy gains in learning ERC and dependency explainability in ERC results.

Methodology

Problem Definition

Given a multi-turn multi-party (or dyadic) dialogue $D = \{u_1, u_2, ..., u_{|D|}\}$, ERC aims to identify emotion labels $Y = \{y_1, y_2, ..., y_{|D|}\}$ for utterance-speaker pairs $\{(u_1, sp_1), (u_2, sp_2), ..., (u_{|D|}, sp_{|D|})\}$. $|D|$ is the number of dialogues.

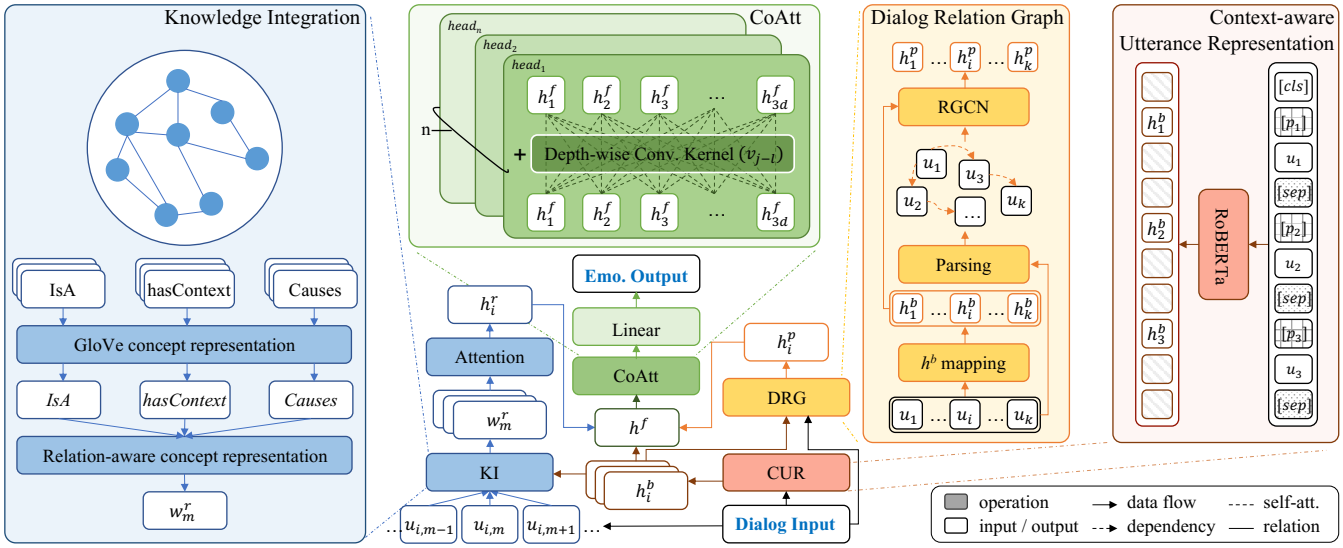

Figure 2: SKIER framework. It contains four main components, i.e., context-aware utterance representation module (CUR), knowledge integration (KI) module, dialogue relation graph (DRG) module and symbolic knowledge fusion module (CoAtt).

Note that the speakers of the ith utterance sp_i and jth utterance sp_j ($i \neq j$) can be the same speaker k and share the same special token $[p_k]$. Here, an utterance in a conversation consists of M tokens, i.e., $u_i = \{u_{i,1}, u_{i,2}, ..., u_{i,M}\}$. Emotion labels are defined by an employed dataset. Taking the MELD dataset as an example, the emotion labels include ANGER, DISGUST, SADNESS, JOY, SURPRISE and FEAR from Ekman's six basic emotions (Ekman 1992), and an additional NEUTRAL class.

Model Overview

Fig. 2 shows the structure of our proposed SKIER. It consists of four technical components. First, a RoBERTa-based context-aware utterance-level representation (CUR) module is used to integrate the speaker dependency and utterance interactions into a single sequence embedding. The generated utterance-level embedding is fed to the later three fusion modules. The second module is DRG construction. DRG utilizes a discourse parser to discover the inter-dependencies (Wang et al. 2021) between utterances and regards the dependency-based dialogue structure as utterance-level symbolic knowledge. We exploit relational graph convolutional networks (RGCN) (Schlichtkrull et al. 2018) to embed the utterance-level symbolic knowledge. The third module is knowledge integration (KI) that leverages a concept-level commonsense knowledge base, ConceptNet (Speer, Chin, and Havasi 2017), and a sentiment lexicon knowledge base, SenticNet (Cambria et al. 2022) to generate the relation-aware concept representation (RACR) of an utterance from the concept-level symbolic knowledge. Finally, a 3-channel convolutional self-attention mechanism (CoAtt) (Dai et al. 2021; Shaw, Uszkoreit, and Vaswani 2018) is applied for fusing the symbolic knowledge. The output is used for affective classification.

Context-Aware Utterance-Level Representation

We integrate speaker information and utterances into a single sequence, and employ a RoBERTa to capture the interactions among utterances and speaker dependencies, simultaneously. Specifically, we first add several special tokens to represent different speakers in a conversation, e.g., $[p_1]$ and $[p_2]$ for a dyadic conversation. Next, all the utterances along with the corresponding speaker tokens are concatenated in a sequence. For instance, a 3-turn dialogue can be represented as $x = \{[\text{cls}], [\text{p}_1], u_1, [\text{sep}], [\text{p}_2], u_2, [\text{sep}], [\text{p}_1], u_3\}$, where special tokens are in between square brackets. The output of RoBERTa-encoded x is $h = RoBERTa(x)$, where $h \in \mathbb{R}^{d \times N}$ and d is the output dimension of the RoBERTa. We obtain the contextual embedding of utterance u_i through $h_i^b := h_j$, where $j : x_{j+1} = u_{i,1}$. This means we first find the index (j) of the last speaker special token before u_i, and then regard the jth vector of h as the utterance-level representation of u_i. Here, h_i^b is the RoBERTa embedding incorporated with context and speaker dependencies.

Dialogue Relation Graph Construction

Previous studies show that dialogue structures are beneficial for several downstream NLP tasks, including dialogue summarization (Chen and Yang 2021) and dialogue comprehension (He, Zhang, and Zhao 2021). Thus, deep learning-based affective classifiers would benefit from integrating DRGs (utterance-level symbolic knowledge).

Following the definition of discourse relations from Asher et al. (2016), we pre-train a dialogue parsing model Deep Sequential (Shi and Huang 2019) on a multi-party dialogue corpus STAC (Asher et al. 2016). We then utilize the pre-trained dialogue parser to parse dialogues in MELD, EmoryNLP and DailyDialog. However, STAC was collected from the game board of an online game *The Settlers of Catan* whose conversation domain and language style are differ-

ent from our ERC datasets (MELD and EmoryNLP sourced from TV shows; DailyDialog sourced from English learning websites). Besides, Liu and Chen (2021) argued that the model trained on the STAC dataset had a very limited generalization ability over the Molweni dataset (Li et al. 2020b) from another domain and vice versa. With TL (Zhuang et al. 2020), a small amount of annotated data from Molweni can improve the generalization ability of the model trained on STAC by a large margin and vice versa (see experiments later). Hence, we invited two expert annotators to manually label discourse graphs of 50 dialogues in MELD, EmoryNLP and DailyDialog, respectively. With these annotated dialogues, we can transfer the knowledge from the pre-trained Deep Sequential model to a new domain, and mitigate its prediction biases (Mao et al. 2022b). Since symbolic knowledge is mostly represented as graphs /knowledge bases (Li, Wang, and Zhu 2020; Narasimhan, Lazebnik, and Schwing 2018), we construct a DRG $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$ for a given parsed conversation, where $\mathcal{V}$ is the set of nodes representing utterances in a conversation; $\mathcal{E}$ is the set of edges between each parent-child node pair; $\mathcal{T}$ is the set of edge types that corresponds to the edges in $\mathcal{E}$. For instance, u_i and u_j are two nodes in a conversation ($i < j$), where $e_{i,j}$ is the edge between parent node i and child node j, and $t_{i,j}$ represents a certain relation type such as *Comment* in Fig 1.

We employ RGCN as the base graph network to encode the DRG, because RGCN naturally supports the calculation of different edge types, e.g., *Comment* is learned differently from *QAP*. RGCN may have multi-layers, where each layer corresponds to a pre-defined directed acyclic graph $\mathcal{G}$. The lth layer of RGCN is given by:

$$g_i^{(l+1)} = \sigma(\sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{N}_i^t} \frac{1}{c_{i,t}} W_t^{(l)} g_j^{(l)} + W_0^{(l)} g_i^{(l)}), \quad (1)$$

where $g_i^{(l)}$ is the hidden state of a child node h_i^b in the lth layer, $g_j^{(l)}$ is that of a parent node h_j^b, and $g_i^{(0)} = h_i^b$. $\mathcal{N}_i^t$ is the set of parent node indices of child node h_i^b in relation $t \in \mathcal{T}$. $c_{i,t}$ is a normalization constant set as default $c_{i,t} = |\mathcal{N}_i^t|$ (Schlichtkrull et al. 2018). $\sigma(\cdot)$ is ReLU (Glorot, Bordes, and Bengio 2011) activation function. Here, we define h_i^b and h_j^b are the inputs of the RGCN model; h_i^p is the output, which represents a dialogue structure-aware embedding of utterance u_i.

Integrating Knowledge Bases in ERC

The aforementioned utterance-level symbolic knowledge depends on discourse. It cannot provide knowledge beyond context. Meanwhile, recent studies showed the effectiveness of external knowledge bases in many NLP tasks (Mao, Lin, and Guerin 2018; Zhong, Wang, and Miao 2019; Ghosal et al. 2020a; Mao et al. 2022a). Hence, we propose to utilize a commonsense knowledge base ConceptNet (Speer, Chin, and Havasi 2017) and a sentiment lexicon knowledge base SenticNet (Cambria et al. 2022) for the ERC task. ConceptNet is a large-scale knowledge graph of concepts. It contains varieties of concepts recorded in triplets, e.g., <*concept1*, relation, *concept2*>.

We define *concept1* as the source node and *concept2* as the destination node. The triplet is an assertion with a confident score[2] (s), e.g., <*alcohol*, Causes, *drunkenness*> with $s = 2$, <*alcohol*, Causes, *amnesia*> with $s = 1$, <*alcohol*, IsA, *addictive*> with $s = 1$. Our goal is to learn the concept representation of *alcohol* under each relation by integrating its different destination nodes, e.g. *drunkenness* and *amnesia* with various s; Then, we generate the RACR by merging the concept representations among different relations. The current version of ConceptNet has around 5.9M assertions, 3.1M concepts and 38 relations. SenticNet contains a large number of words with sentiment intensity scores, ranging from -1 to 1, which measures the sentiment intensities of both positive and negative words.

Concept Representation Three main relations, e.g., *IsA*, *HasContext*, and *Causes* are used out of 38 ConceptNet relations. This is because we assume the concepts under the three relations are prone to containing sentiment. For each source node $u_{i,m}$ in u_i and each relation r_j in the three ConceptNet relations ($j \in \{1, 2, 3\}$), we retrieve all their destination nodes with confidence scores more than 1. As a result, we have three sets of triplets for the source node $u_{i,m}$: $\{(u_{i,m}, r_j, o_{j,k})\}_j$, where $o_{j,k}$ denotes the kth destination node of the source node $u_{i,m}$ in relation type r_j, $k \in \{1, 2, ..., N_d\}$, and N_d is the total number of destination nodes. In addition, we also have confidence scores for each triplet. GloVe (Pennington, Socher, and Manning 2014) is used to generate word embeddings for concept tokens. To enrich utterance embeddings with symbolic concept knowledge, we compute the concept representation for each source node $u_{i,m}$ in u_i by taking triplet relations into account. The concept representation $\mathbf{c}_{m,j} \in \mathbb{R}^{d \times 1}$ for $u_{i,m}$ is given by:

$$\mathbf{c}_{m,j} = \sum_{k=1}^{N_d} \alpha_k \cdot \mathbf{o}_{j,k}, \quad (2)$$

where $\mathbf{o}_{j,k} \in \mathbb{R}^{d \times 1}$ is the embedding of token $o_{j,k}$. α_k denotes the corresponding attention weight which is given by:

$$\alpha_k = \textit{softmax}(\omega_k), \quad (3)$$

where ω_k is the calculated weight for $o_{j,k}$. The calculation of weight ω_k is of vital importance, as it measures the contribution of the destination node $o_{j,k}$ towards $u_{i,m}$ in terms of enriching the concept representation of $u_{i,m}$. Motivated by the assumption that important concepts are semantically relevant to dialogue context and have strong sentiment intensities (Zhong, Wang, and Miao 2019), we compute ω_k by measuring the context relatedness (ω_k^c) and the affective intensity (ω_k^a) of the destination node $o_{j,k}$:

$$\omega_k^c = min{-}max(s_k) \cdot |\cos(h_i^b, \mathbf{o}_{j,k})|, \quad (4)$$

where s_k is the confidence score; *min-max(·)* is a min-max scaling function; *cos(·)* is a cosine similarity function; $h_i^b \in$

[2]The confident scores range from 0.1 to 22 (Chen et al. 2019). Confidence larger than 1 is considered a confident fact, according to (Zhong, Wang, and Miao 2019).

$\mathbb{R}^d$ is the context-aware party-dependent representation of the ith utterance given by the RoBERTa. ω_k^a is given by:

$$\omega_k^a = min{-}max(sentic(o_{j,k})), \tag{5}$$

where *sentic*($o_{j,k}$) is the sentiment intensity score of the destination node $o_{j,k}$ from SenticNet. Then, ω_k is given by:

$$\omega_k = \lambda_k \cdot \omega_k^c + (1 - \lambda_k) \cdot \omega_k^a, \tag{6}$$

where λ_k is a hyperparameter.

Relation-Aware Concept Representation With the aforementioned equations, we obtain three concept representations of $u_{i,m}$, namely $\mathbf{c}_{m,1}$, $\mathbf{c}_{m,2}$, $\mathbf{c}_{m,3}$. Motivated by the contextualized entity learning of Qiao et al. (2020), we calculate the RACR of $u_{i,m}$ by:

$$\mathbf{w}_m^r = \mathbf{w}_m + \sum_{(r_j, o_{j,k}) \in C_m} \beta_{j,k} \cdot (\mathbf{r}_j \odot \mathbf{c}_{m,j}), \tag{7}$$

where $\mathbf{w}_m$ is the GloVe embedding of $u_{i,m}$; $\mathbf{r}_j$ is the randomly initialized relation embedding of r_j; $\beta_{j,k} = \frac{\exp(q_{j,k})}{\sum_{(r_{j'}, o_{j'}, k')} \exp(q_{j',k'})}$ represents the importance of each concept representation to $u_{i,m}$; $\odot$ is an element-wise multiplication. The concept context C_m of $u_{i,m}$ is defined as a set of $(r_j, o_{j,k})$ pairs. The $q_{j,k}$ represents the score for each possible triplet $(u_{i,m}, r_j, o_{j,k})$, which is calculated via the score function with DistMult (Yang et al. 2014):

$$q_{j,k} = \mathbf{w}_m^\top (\mathbf{r}_j \odot \mathbf{o}_{j,k}). \tag{8}$$

We then apply a dot-product attention (Vaswani et al. 2017) to convert the word-level concept representations $\mathbf{w}_m^r$ into an utterance-level RACR h_i^r. The attention weight $\gamma_{i,m}$ is obtained by measuring the relevance between contextual embedding h_i^b and $\mathbf{w}_m^r$.

Symbolic Knowledge Fusion

We have obtained structure-aware knowledge h_i^p, relation-aware concept knowledge h_i^r, and context-aware representation h_i^b, incorporated with speaker-dependency. Next, we introduce CoAtt to fuse the symbolic knowledge h_i^p and h_i^r into the contextual embedding h_i^b for emotion recognition.

Convolutional Self-Attention Fusion CoAtt was originally proposed for computer vision (Dai et al. 2021), while we are the first to apply it in NLP. CoAtt is supposed to combine the advantages of both convolution and self-attention. We feed the aforementioned symbolic knowledge features h_i^r, h_i^p, and context feature h_i^b, into a multi-head CoAtt. We firstly obtain $h^f \in \mathbb{R}^{d \times 3}$ via the concatenation operation in Eq. (9) Then we use a CoAtt to capture the interactions among the features and generate $x^{(k)} \in \mathbb{R}^{d \times d_h}$ in each head. The jth element of $x^{(k)}$ ($x_j^{(k)} \in \mathbb{R}^{1 \times d_h}$) is computed as Eq. (11).

$$h^f = [h_i^b \oplus h_i^r \oplus h_i^p] \tag{9}$$

$$(Q, K, V) = (h^f W_Q, h^f W_K, h^f W_V) \tag{10}$$

$$x_j^{(k)} = \sum_{l \in \mathcal{I}} \frac{\exp(Q_j K_l^\top + v_{j-l})}{\sum_{l' \in \mathcal{I}} \exp(Q_j K_{l'}^\top + v_{j-l'})} V_j \tag{11}$$

Dataset		Train	Dev	Test	Label	Metrics
MELD	u	9989	1109	2610	7/3	Weighted Avg F1
	d	1038	114	280		
EmoryNLP	u	9934	1344	1328	7/3	Weighted Avg F1
	d	713	99	85		
DailyDialog	u	87170	8069	7740	7(6)	Macro & Micro F1
	d	11118	1000	1000		

Table 1: Statistical information for the datasets (u and d refer to utterance and dialogue). MELD and EmoryNLP both have 7 emotion and 3 sentiment labels, and we use weighted avg F1 as the evaluation metric. For DailyDialog dataset, we use 7 emotion labels in training and measure Micro-F1 for only 6 emotion labels excluding NEUTRAL.

$\oplus$ is a concatenation operation; $W_Q, W_K, W_V \in \mathbb{R}^{3 \times d_h}$; d_h is the dimension of head; v_{j-l} is a scalar bias between Q_j and K_l; $\mathcal{I} \in [0, d) \cap \mathbb{N}$. Next, the outputs from n heads are concatenated and projected into the final output $x \in \mathbb{R}^{d \times 1}$ through a linear layer $W_O \in \mathbb{R}^{n d_h \times 1}$ as Eq. (12). Finally, x is connected with a fully connected layer for classification.

$$x = [x^{(1)} \oplus x^{(2)} \oplus \cdots \oplus x^{(n)}] W_O \tag{12}$$

We choose cross entropy as the loss function and utilize L2-regularization to alleviate overfitting. The loss (L) is:

$$L = -\frac{1}{\sum_{j=1}^{|D|} N_j} \sum_{j=1}^{|D|} \sum_{i=1}^{N_j} \log P_{i,j}\left[y_{i,j}\right] + \rho \|\theta\|_2, \tag{13}$$

where N_j is the number of utterances in the jth conversation; $P_{i,j}$ is the probability distribution of label $y_{i,j}$ for the ith utterance in the jth conversation; ρ is the L2-regularization weight; θ is trainable parameters.

Experiment

Datasets

We employed three public datasets for benchmarking (Table 1). **DailyDialog** (Li et al. 2017) derives from human daily communication. The data were sourced from English learning websites. The emotion labels include Ekman's six basic emotions and a NEUTRAL class. **MELD** (Poria et al. 2019) contains TV show scripts, collected from *Friends*. The utterances involve multiple parties. The emotion labels are also from Ekman's six basic emotions plus a NEUTRAL class. Sentiment labels {POSITIVE, NEGATIVE, NEUTRAL} are also provided in this dataset. We use the textual data in the dataset. **EmoryNLP** (Zahiri and Choi 2018) is a multi-party ERC dataset, sourced from *Friends* TV show scripts. The emotion labels are {JOYFUL, PEACEFUL, POWERFUL, SCARED, MAD, SAD, NEUTRAL}. Sentiment labels were not provided but can be categorized by neutral:{NEUTRAL}, positive:{JOYFUL, POWERFUL, PEACEFUL}, negative:{SCARED, SAD, MAD}.

Baselines

CNN (Kim 2014) is a convolutional neural network model for sentence classification.

KET (Zhong, Wang, and Miao 2019) employs a knowledge-enriched Transformer, incorporating lexicon-level ConceptNet and sentiment knowledge to enhance ERC.

DialogueGCN (DiGCN) (Ghosal et al. 2020b) learns the intra- and inter-speaker dependencies via GCN. The input features are 300-dimensional GloVe embeddings.

DialogueRNN(DiRNN) (Majumder et al. 2019) exploits three groups of GRUs to represent the speaker states, context, and emotion, respectively. Ghosal et al. shows the performance of DiRNN (**RoDiRNN**) based on a RoBERTa.

COSMIC (Ghosal et al. 2020a) introduces commonsense knowledge, such as mental states and causal relations to support ERC. GRUs are used to encode the knowledge.

DialogXL(DiXL) (Shen et al. 2021a) proposes dialog-aware self-attention to learn intra- and inter-speaker dependencies.

DAG (Shen et al. 2021b) utilizes a directed acyclic graph to encode the intrinsic structure within a dialogue.

P-CKG (Li et al. 2021) considers the psychological interactions between utterances and proposes a commonsense knowledge enhanced graph transformer model.

T-GCN (Lee and Choi 2021) treats the ERC as dialogue-based relation extraction and designs a GCN-based model, learning the way people understand dialogues.

CoMPM (Lee and Lee 2022) extracts external knowledge using a RoBERTa and integrates the speaker's pre-trained memory into the context model to improve ERC results.

Setups

We used RoBERTa-Large from HuggingFace[3]. The optimizer was AdamW (Loshchilov and Hutter 2018) with an initial learning rate of 1e-5. We used a linear scheduler during training. The maximum value of 5 was used for the gradient clipping. The actual number of dialogue relations was set to $\{9, 10, 11\}$ for EmoryNLP, MELD and DailyDialog, respectively, because some dialogue relations, e.g., background and narration, do not exist in the parsed datasets. The batch size was 1. The dropout rate was 0.2. λ_k in Eq. 6 was 0.5. The number of destination nodes was 3. All experiments were conducted on a V100 GPU with 16 GB memory. We reported the average score of 3 random runs on test sets.

Results

Dialogue Parsing Analysis

As mentioned, the dialogue parsing model has poor generalization ability in a new domain. Hence, we conducted TL with annotated seed samples. Experiments were conducted on STAC and Molweni datasets to investigate the effectiveness of the TL mechanism. The results in Table 2 show the performance gains of a small number of annotated samples on a cross-domain dataset.

ERC Result

Table 3 shows the results of the baselines and SKIER. The baselines are categorized by methods based on GloVe and other pre-trained language models (PM). To demonstrate the effectiveness of CoAtt, we introduce its competitor solutions, e.g., SKIER-l (the symbolic knowledge is fused by a fully connected layer), and SKIER-a (the knowledge is fused by an element-wise addition). As seen in Table 3, SKIER surpasses the strongest baseline on each metric by 1.74% on average. For example, SKIER outperforms the strongest baseline (CoMPM) by 1.97% and 0.87% on sentiment analysis (3-cls) and emotion detection (7-cls) setups on MELD dataset, although CoMPM has more than twice

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

Mode	**S to M**		**M to S**	
Metrics	**F1_bi**	**F1_mul**	**F1_bi**	**F1_mul**
No Transfer	54.83	32.78	44.47	9.61
Transfer(10)	72.12	47.54	66.17	37.46
Transfer(50)	73.41	50.56	68.00	43.22
Transfer(100)	75.28	52.42	68.92	45.56

Table 2: TL analysis. The evaluation metrics are the f1 score of binary link prediction (F1_bi) and multi-class relation prediction (F1_mul) in dialogue parsing. *S to M* means training on the STAC and testing on the Molweni dataset, and vice versa. The values in brackets are annotated seed samples.

	Methods	**MELD**		**EmoryNLP**		**DailyDialog**	
		Weighted Avg F1				Macro	Micro
		3-cls	7-cls	3-cls	7-cls		
GloVe-based	CNN	64.25	55.02	38.05	32.59	36.87	50.32
	DiGCN	-	58.37	-	34.29	49.95	53.73
	KET	-	58.18	-	34.39	-	53.37
	DiXL	-	62.41	-	34.73	-	54.93
	DiRNN	66.10	57.03	48.93	31.70	41.80	55.95
PM-based	COSMIC	73.20	65.21	56.51	38.11	51.05	58.48
	DAG	-	63.65	-	39.02	-	59.33
	P-CKG	-	65.18	-	38.80	51.59	59.75
	T-GCN		65.36	-	39.24	-	**61.91**
	RoDiRNN	72.12	62.02	55.28	37.29	48.20	55.16
	RoBERTa	72.14	63.61	55.36	37.44	49.65	57.32
	CoMPM	73.08	66.52	57.14	37.37	53.15	60.34
	SKIER	**75.05**	**67.39**	**60.08**	**40.07**	**56.68**	**62.31**
	SKIER-l	**74.73**	**66.99**	57.98	**39.49**	**56.39**	61.72
	SKIER-a	74.17	66.91	**59.39**	39.53	54.02	61.03

Table 3: Performance comparisons on three benchmark datasets. The top 2 best results are in bold.

Component	**MELD**	**EmoryNLP**	**DailyDialog**
SKIER	67.39	40.07	56.68(Macro)
w/o DRG	$65.27_{\downarrow 2.12}$	$38.54_{\downarrow 1.53}$	$55.70_{\downarrow 0.98}$
w/o KI	$66.10_{\downarrow 1.29}$	$38.56_{\downarrow 1.51}$	$52.54_{\downarrow 4.14}$
w/o DRG & KI	$64.08_{\downarrow 3.31}$	$38.10_{\downarrow 1.97}$	$49.73_{\downarrow 6.95}$
w/o TL	$65.87_{\downarrow 1.52}$	$39.50_{\downarrow 0.57}$	$53.38_{\downarrow 3.30}$
w/o DRS	$65.73_{\downarrow 1.66}$	$39.25_{\downarrow 0.82}$	$56.33_{\downarrow 0.35}$
w/o PDR	$66.24_{\downarrow 1.15}$	$39.08_{\downarrow 0.99}$	$56.06_{\downarrow 0.62}$
w/o RACR	$65.80_{\downarrow 1.59}$	$39.14_{\downarrow 0.93}$	$54.99_{\downarrow 1.69}$
w/o IsA	$65.95_{\downarrow 1.44}$	$38.79_{\downarrow 1.28}$	$55.69_{\downarrow 0.99}$
w/o HasContext	$65.97_{\downarrow 1.42}$	$38.44_{\downarrow 1.63}$	$55.68_{\downarrow 1.00}$
w/o Causes	$66.36_{\downarrow 1.03}$	$38.55_{\downarrow 1.52}$	$55.21_{\downarrow 1.47}$

Table 4: Ablation studies on three datasets.

the parameters of SKIER. Many MELD data are short conversations with multiple speakers, highlighting the significance of capturing utterance dependencies. SKIER incorporates dependencies via DRG, and thus yields better results. The overall performance on EmoryNLP is worse than that on MELD, as many utterances are not grammatically complete and contain almost no emotion-specific words. Nevertheless, SKIER largely improves the performance on the sentiment classification task by incorporating commonsense knowledge. SKIER significantly improves the state-of-the-art performance by 3.53% in Macro-F1. SKIER surpasses SKIER-l and SKIER-a on the three datasets, showing that the CoAtt module is effective for fusing external knowledge and DRGs, as it captures interactions among each dimension and channel.

Ablation Study

We conducted ablation studies to investigate the utilities of the key components of SKIER. As shown in Table 4, DRG and KI modules are crucial to SKIER. When we removed DRG from SKIER, w/o DRG performance dropped, e.g., from 67.39% to 65.27% on MELD. The weighted average F1 score decreased from 67.39% to 66.10%, if we kept the DRG module and disabled the KI module (w/o KI). After removing both components (w/o DRG & KI), the remaining part equaled a RoBERTa classifier. Its F1 score further declined. Without transfer learning (w/o TL), the performance dropped by 1.52% on MELD. As a portion of dialogue relations do not exist in the parsed datasets, we simplified the number of relations. The ablation result indicated that SKIER benefited from the dialogue relation simplification (DRS), because the w/o DRS model is weaker.

If we removed the parsed dialogue relations (PDR) and simply used the linking information, there is a loss in the w/o PDR model. The result proved that a complete DRG is indispensable, as it provides necessary fine-grained utterance dependency information. We proposed a relation-aware concept representation (RACR) mechanism, taking different relations of concepts into account. The effect of RACR can be confirmed by comparing the performances of SKIER and w/o RACR. Moreover, we studied the influence of the three selected relations from ConceptNet. When removing one of the relations, we observed a significant drop in performance on both datasets.

Hyperparameter Analysis

We analyzed the influence of the number of destination nodes in this section. By viewing Fig. 3, we observed a trend that the model achieved the best results on both datasets by using 3 destination nodes. Using more nodes improved the computing costs, whereas it did not yield accuracy gains. Thus, we set the number of destination nodes as 3.

Case Study

We illustrated a case study on a conversation snippet of MELD test set between A and B. In Fig. 4, we observed that the utterance indexed by A_1 contains 1 positive and no negative emotional word. However, it received the transmitted SADNESS emotion from (<hurt, *Causes*, ache>) in

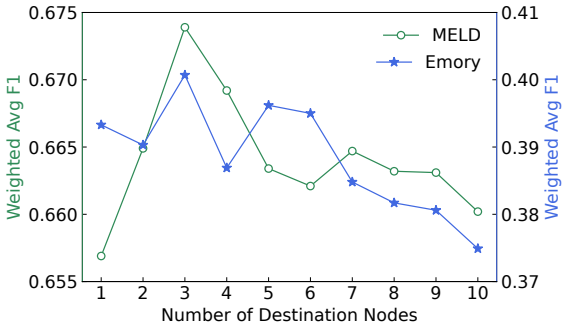

Figure 3: The number of destination nodes analysis.

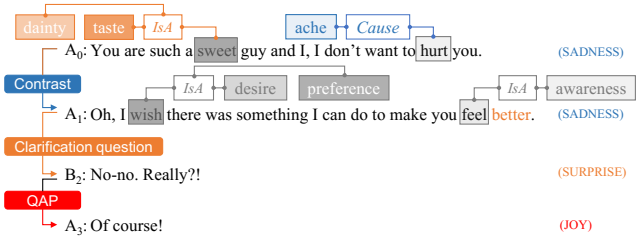

Figure 4: Case study. In the KI module, the destination node in darker color obtained a higher attention score compared with other destination nodes, e.g., *taste* > *dainty*; The source node in darker gray gets a higher weight γ than other source nodes, e.g., *wish* > *feel*.

utterance A_0 through PDR *Contrast*. In addition, the original emotion of the utterance indexed by B_2 is ambiguous as it does not have an emotion-specific word. Nevertheless, it got the positive word "better" transmitted directly via relation *Clarification question*, and (<sweet, *IsA*, taste>) & (<sweet, *IsA*, dainty>) in utterance A_0 indirectly via relations *Contrast* & *Clarification question*. This enabled our SKIER to recognize the SURPRISE emotion in utterance B_2. In short, the case indicated that DRG and KI modules allowed SKIER to explore the informative words or structures under the iceberg and exploit the symbolic knowledge to improve the emotion (sentiment) classification accuracy. Moreover, the predicted dependency relations also explain the emotion predictions with linguistic intuition.

Conclusion

In this paper, we proposed a neurosymbolic model for ERC named SKIER. The model explicitly integrated dialogue structure knowledge and commonsense knowledge. To effectively fuse the multiple-level symbolic knowledge, SKIER included relational graph convolutional network, relation-aware concept representation, and convolutional self-attention techniques, yielding state-of-the-art performances on three ERC datasets. Since there is a big room for improving dialogue dependency parser performance, we will study this in future work.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

References

Asher, N.; Hunter, J.; Morey, M.; Benamara, F.; and Afantenos, S. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2721–2727.

Cambria, E.; Liu, Q.; Decherchi, S.; Xing, F.; ; and Kwok, K. 2022. SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis. In *LREC*, 3829–3839.

Chen, J.; and Yang, D. 2021. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1380–1391.

Chen, T.; Xu, R.; He, Y.; and Wang, X. 2017. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72: 221–230.

Chen, X.; Chen, M.; Shi, W.; Sun, Y.; and Zaniolo, C. 2019. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3363–3370.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.

Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34: 3965–3977.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200.

Ge, M.; Mao, R.; and Cambria, E. 2022. Explainable Metaphor Identification Inspired by Conceptual Metaphor Theory. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 10681–10689.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020a. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2470–2481.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2020b. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323. JMLR Workshop and Conference Proceedings.

He, Y.; Zhang, Z.; and Zhao, H. 2021. Multi-tasking Dialogue Comprehension with Discourse Parsing. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, 69–79.

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

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Doha, Qatar: Association for Computational Linguistics.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR*. OpenReview.net.

Lee, B.; and Choi, Y. S. 2021. Graph Based Network with Contextualized Representations of Turns in Dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 443–455.

Lee, J.; and Lee, W. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. 5669–5679.

Li, G.; Wang, X.; and Zhu, W. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1227–1235.

Li, J.; Ji, D.; Li, F.; Zhang, M.; and Liu, Y. 2020a. HiTrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4190–4200.

Li, J.; Lin, Z.; Fu, P.; and Wang, W. 2021. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1204–1214. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, J.; Liu, M.; Kan, M.-Y.; Zheng, Z.; Wang, Z.; Lei, W.; Liu, T.; and Qin, B. 2020b. Molweni: A Challenge Multiparty Dialogues-based Machine Reading Comprehension Dataset with Discourse Structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2642–2652.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 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)*, 986–995.

Liang, B.; Su, H.; Gui, L.; Cambria, E.; and Xu, R. 2022. Aspect-Based Sentiment Analysis via Affective Knowledge Enhanced Graph Convolutional Networks. *Knowledge-Based Systems*, 235(107643).

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; and Chen, N. F. 2021. Improving Multi-Party Dialogue Discourse Parsing via Domain Integration. *arXiv preprint arXiv:2110.04526*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ma, Y.; Nguyen, K. L.; Xing, F. Z.; and Cambria, E. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64: 50–70.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6818–6825.

Mao, R.; and Li, X. 2021. Bridging Towers of Multitask Learning with a Gating Mechanism for Aspect-based Sentiment Analysis and Sequential Metaphor Identification. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 13534–13542.

Mao, R.; Li, X.; Ge, M.; and Cambria, E. 2022a. MetaPro: A computational metaphor processing model for text pre-processing. *Information Fusion*, 86: 30–43.

Mao, R.; Lin, C.; and Guerin, F. 2018. Word Embedding and WordNet Based Metaphor Identification and Interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, 1222–1231.

Mao, R.; Liu, Q.; He, K.; Li, W.; and Cambria, E. 2022b. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 174–184.

Narasimhan, M.; Lazebnik, S.; and Schwing, A. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in neural information processing systems*, 31.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536.

Qiao, Z.; Ning, Z.; Du, Y.; and Zhou, Y. 2020. Context-Enhanced Entity and Relation Embedding for Knowledge Graph Completion. *arXiv preprint arXiv:2012.07011*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Berg, R. v. d.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, 593–607. Springer.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 464–468.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021a. DialogXL: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13789–13797.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021b. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 1551–1560.

Shi, Z.; and Huang, M. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7007–7014.

Song, X.; Zang, L.; Zhang, R.; Hu, S.; and Huang, L. 2022. Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8542–8546. IEEE.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

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 *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, A.; Song, L.; Jiang, H.; Lai, S.; Yao, J.; Zhang, M.; and Su, J. 2021. A Structure Self-Aware Model for Discourse Parsing on Multi-Party Dialogues. In *IJCAI*, 3943–3949.

Xu, J.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; and Zhou, G. 2019. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *IJCAI*, 5415–5421.

Zhang, D.; Zhang, W.; Li, S.; Zhu, Q.; and Zhou, G. 2020. Modeling both intra-and inter-modal influence for real-time emotion detection in conversations. In *Proceedings of the 28th ACM International Conference on Multimedia*, 503–511.

Zhang, K.; Li, Y.; Wang, J.; Cambria, E.; and Li, X. 2021. Real-Time Video Emotion Recognition based on Reinforcement Learning and Domain Knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1034–1047.

Zhao, Z.; Rao, G.; and Feng, Z. 2017. DFDS: A Domain-Independent Framework for Document-Level Sentiment Analysis Based on RST. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, 297–310. Springer.

Zhong, P.; Wang, D.; and Miao, C. 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*, 165–176.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.