Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations

Peixiang Zhong[1,2], **Di Wang**[1], **Chunyan Miao**[1,2,3]
[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly
[2]Alibaba-NTU Singapore Joint Research Institute
[3]School of Computer Science and Engineering
Nanyang Technological University, Singapore
peixiang001@e.ntu.edu.sg, {wangdi, ascymiao}@ntu.edu.sg

Abstract

Messages in human conversations inherently convey emotions. The task of detecting emotions in textual conversations leads to a wide range of applications such as opinion mining in social networks. However, enabling machines to analyze emotions in conversations is challenging, partly because humans often rely on the context and commonsense knowledge to express emotions. In this paper, we address these challenges by proposing a Knowledge-Enriched Transformer (KET), where contextual utterances are interpreted using hierarchical self-attention and external commonsense knowledge is dynamically leveraged using a context-aware affective graph attention mechanism. Experiments on multiple textual conversation datasets demonstrate that both context and commonsense knowledge are consistently beneficial to the emotion detection performance. In addition, the experimental results show that our KET model outperforms the state-of-the-art models on most of the tested datasets in F1 score.

1 Introduction

Emotions are "generated states in humans that reflect evaluative judgments of the environment, the self and other social agents" (Hudlicka, 2011). Messages in human communications inherently convey emotions. With the prevalence of social media platforms such as Facebook Messenger, as well as conversational agents such as Amazon Alexa, there is an emerging need for machines to understand human emotions in natural conversations. This work addresses the task of detecting emotions (e.g., happy, sad, angry, etc.) in textual conversations, where the emotion of an utterance is detected in the conversational context. Being able to effectively detect emotions in conversations leads to a wide range of applications ranging from opinion mining in social media platforms (Chatterjee et al., 2019) to building emotion-aware conversational agents (Zhou et al., 2018a).

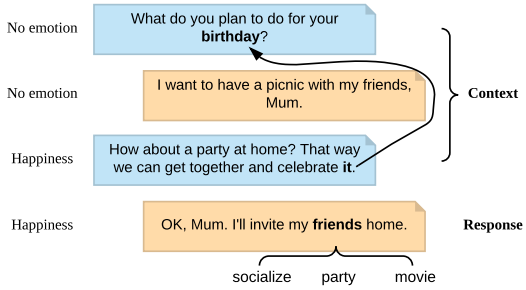

Figure 1: An example conversation with annotated labels from the DailyDialog dataset (Li et al., 2017). By referring to the context, "it" in the third utterance is linked to "birthday" in the first utterance. By leveraging an external knowledge base, the meaning of "friends" in the forth utterance is enriched by associated knowledge entities, namely "socialize", "party", and "movie". Thus, the implicit "happiness" emotion in the fourth utterance can be inferred more easily via its enriched meaning.

However, enabling machines to analyze emotions in human conversations is challenging, partly because humans often rely on the context and commonsense knowledge to express emotions, which is difficult to be captured by machines. Figure 1 shows an example conversation demonstrating the importance of context and commonsense knowledge in understanding conversations and detecting implicit emotions.

There are several recent studies that model contextual information to detect emotions in conversations. Poria et al. (2017) and Majumder et al. (2019) leveraged recurrent neural networks (RNN) to model the contextual utterances in sequence, where each utterance is represented by a feature vector extracted by convolutional neural networks (CNN) at an earlier stage. Similarly, Hazarika et al. (2018a,b) proposed to use extracted CNN

features in memory networks to model contextual utterances. However, these methods require separate feature extraction and tuning, which may not be ideal for real-time applications. In addition, to the best of our knowledge, no attempts have been made in the literature to incorporate commonsense knowledge from external knowledge bases to detect emotions in textual conversations. Commonsense knowledge is fundamental to understanding conversations and generating appropriate responses (Zhou et al., 2018b).

To this end, we propose a Knowledge-Enriched Transformer (KET) to effectively incorporate contextual information and external knowledge bases to address the aforementioned challenges. The Transformer (Vaswani et al., 2017) has been shown to be a powerful representation learning model in many NLP tasks such as machine translation (Vaswani et al., 2017) and language understanding (Devlin et al., 2018). The self-attention (Cheng et al., 2016) and cross-attention (Bahdanau et al., 2014) modules in the Transformer capture the intra-sentence and inter-sentence correlations, respectively. The shorter path of information flow in these two modules compared to gated RNNs and CNNs allows KET to model contextual information more efficiently. In addition, we propose a hierarchical self-attention mechanism allowing KET to model the hierarchical structure of conversations. Our model separates context and response into the encoder and decoder, respectively, which is different from other Transformer-based models, e.g., BERT (Devlin et al., 2018), which directly concatenate context and response, and then train language models using only the encoder part.

Moreover, to exploit commonsense knowledge, we leverage external knowledge bases to facilitate the understanding of each word in the utterances by referring to related knowledge entities. The referring process is dynamic and balances between relatedness and affectiveness of the retrieved knowledge entities using a context-aware affective graph attention mechanism.

In summary, our contributions are as follows:

- For the first time, we apply the Transformer to analyze conversations and detect emotions. Our hierarchical self-attention and cross-attention modules allow our model to exploit contextual information more efficiently than existing gated RNNs and CNNs.
- We derive dynamic, context-aware, and emotion-related commonsense knowledge from external knowledge bases and emotion lexicons to facilitate the emotion detection in conversations.
- We conduct extensive experiments demonstrating that both contextual information and commonsense knowledge are beneficial to the emotion detection performance. In addition, our proposed KET model outperforms the state-of-the-art models on most of the tested datasets across different domains.

2 Related Work

Emotion Detection in Conversations: Early studies on emotion detection in conversations focus on call center dialogs using lexicon-based methods and audio features (Lee and Narayanan, 2005; Devillers and Vidrascu, 2006). Devillers et al. (2002) annotated and detected emotions in call center dialogs using unigram topic modelling. In recent years, there is an emerging research trend on emotion detection in conversational videos and multi-turn Tweets using deep learning methods (Hazarika et al., 2018b,a; Zahiri and Choi, 2018; Chatterjee et al., 2019; Zhong and Miao, 2019; Poria et al., 2019). Poria et al. (2017) proposed a long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) based model to capture contextual information for sentiment analysis in user-generated videos. Majumder et al. (2019) proposed the DialogueRNN model that uses three gated recurrent units (GRU) (Cho et al., 2014) to model the speaker, the context from the preceding utterances, and the emotions of the preceding utterances, respectively. They achieved the state-of-the-art performance on several conversational video datasets.

Knowledge Base in Conversations: Recently there is a growing number of studies on incorporating knowledge base in generative conversation systems, such as open-domain dialogue systems (Han et al., 2015; Asghar et al., 2018; Ghazvininejad et al., 2018; Young et al., 2018; Parthasarathi and Pineau, 2018; Liu et al., 2018; Moghe et al., 2018; Dinan et al., 2019; Zhong et al., 2019), task-oriented dialogue systems (Madotto et al., 2018; Wu et al., 2019; He et al., 2019) and question answering systems (Kiddon et al., 2016; Hao et al., 2017; Sun et al., 2018; Mihaylov and Frank, 2018). Zhou et al. (2018b) adopted structured

knowledge graphs to enrich the interpretation of input sentences and help generate knowledge-aware responses using graph attentions. The graph attention in the knowledge interpreter (Zhou et al., 2018b) is static and only related to the recognized entity of interest. By contrast, our graph attention mechanism is dynamic and selects context-aware knowledge entities that balances between relatedness and affectiveness.

Emotion Detection in Text: There is a trend moving from traditional machine learning methods (Pang et al., 2002; Wang and Manning, 2012; Seyeditabari et al., 2018) to deep learning methods (Abdul-Mageed and Ungar, 2017; Zhang et al., 2018b) for emotion detection in text. Khanpour and Caragea (2018) investigated the emotion detection from health-related posts in online health communities using both deep learning features and lexicon-based features.

Incorporating Knowledge in Sentiment Analysis: Traditional lexicon-based methods detect emotions or sentiments from a piece of text based on the emotions or sentiments of words or phrases that compose it (Hu et al., 2009; Taboada et al., 2011; Bandhakavi et al., 2017). Few studies investigated the usage of knowledge bases in deep learning methods. Kumar et al. (2018) proposed to use knowledge from WordNet (Fellbaum, 2012) to enrich the text representations produced by LSTM and obtained improved performance.

Transformer: The Transformer has been applied to many NLP tasks due to its rich representation and fast computation, e.g., document machine translation (Zhang et al., 2018a), response matching in dialogue system (Zhou et al., 2018c), language modelling (Dai et al., 2019) and understanding (Radford et al., 2018). A very recent work (Rik Koncel-Kedziorski and Hajishirzi, 2019) extends the Transformer to graph inputs and propose a model for graph-to-text generation.

3 Our Proposed KET Model

In this section we present the task definition and our proposed KET model.

3.1 Task Definition

Let $\{X_j^i, Y_j^i\}, i = 1, ...N, j = 1, ...N_i$ be a collection of $\{$*utterance*, *label*$\}$ pairs in a given dialogue dataset, where N denotes the number of conversations and N_i denotes the number of utterances in the ith conversation. The objective of the task is to maximize the following function:

$$\Phi = \prod_{i=1}^{N} \prod_{j=1}^{N_i} p(Y_j^i | X_j^i, X_{j-1}^i, ..., X_1^i; \theta), \quad (1)$$

where $X_{j-1}^i, ..., X_1^i$ denote contextual utterances and θ denotes the model parameters we want to optimize.

We limit the number of contextual utterances to M. Discarding early contextual utterances may cause information loss, but this loss is negligible because they only contribute the least amount of information (Su et al., 2018). This phenomenon can be further observed in our model analysis regarding context length (see Section 5.2). Similar to (Poria et al., 2017), we clip and pad each utterance X_j^i to a fixed m number of tokens. The overall architecture of our KET model is illustrated in Figure 2.

3.2 Knowledge Retrieval

We use a commonsense knowledge base ConceptNet (Speer et al., 2017) and an emotion lexicon NRC_VAD (Mohammad, 2018a) as knowledge sources in our model.

ConceptNet is a large-scale multilingual semantic graph that describes general human knowledge in natural language. The nodes in ConceptNet are concepts and the edges are relations. Each $\langle$*concept1*, *relation*, *concept2*$\rangle$ triplet is an assertion. Each assertion is associated with a confidence score. An example assertion is $\langle$*friends*, *CausesDesire*, *socialize*$\rangle$ with confidence score of 3.46. Usually assertion confidence scores are in the $[1, 10]$ interval. Currently, for English, ConceptNet comprises 5.9M assertions, 3.1M concepts and 38 relations.

NRC_VAD is a list of English words and their VAD scores, i.e., valence (negative-positive), arousal (calm-excited), and dominance (submissive-dominant) scores in the $[0, 1]$ interval. The VAD measure of emotion is culture-independent and widely adopted in Psychology (Mehrabian, 1996). Currently NRC_VAD comprises around 20K words.

In general, for each non-stopword token t in X_j^i, we retrieve a connected knowledge graph $g(t)$ comprising its immediate neighbors from ConceptNet. For each $g(t)$, we remove concepts that are stopwords or not in our vocabulary. We further remove concepts with confidence scores less

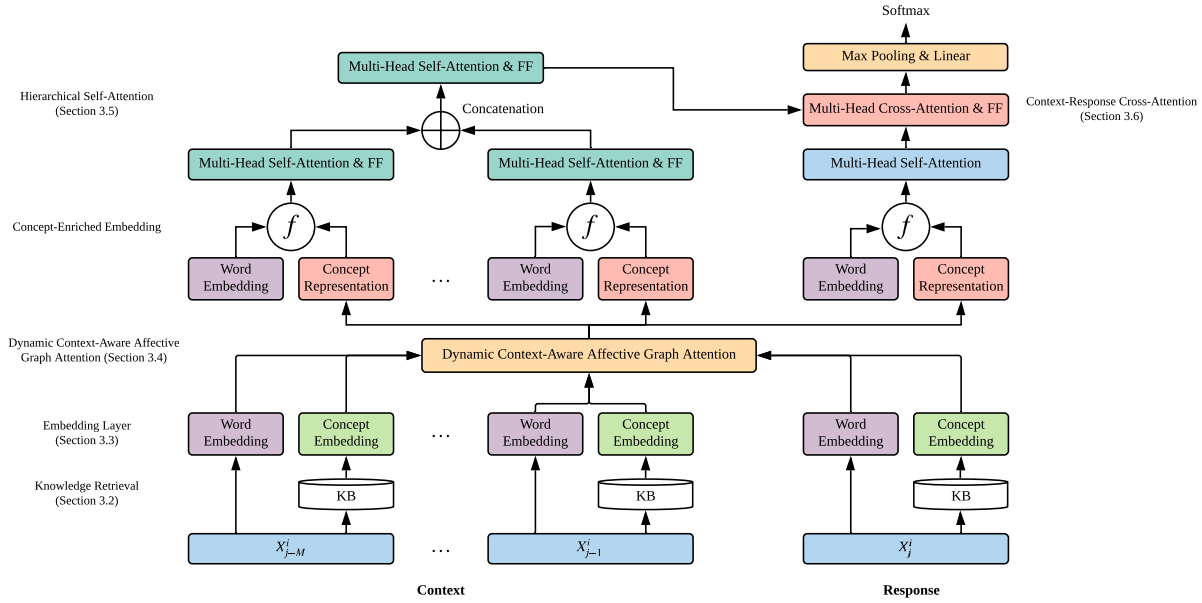

Figure 2: Overall architecture of our proposed KET model. The positional encoding, residual connection, and layer normalization are omitted in the illustration for brevity.

than 1 to reduce annotation noises. For each concept, we retrieve its VAD values from NRC_VAD. The final knowledge representation for each token t is a list of tuples: $(c_1, s_1, VAD(c_1))$, $(c_2, s_2, VAD(c_2))$, ..., $(c_{|g(t)|}, s_{|g(t)|}, VAD(c_{|g(t)|}))$, where $c_k \in g(t)$ denotes the kth connected concept, s_k denotes the associated confidence score, and $VAD(c_k)$ denotes the VAD values of c_k. The treatment for tokens that are not associated with any concept and concepts that are not included in NRC_VAD are discussed in Section 3.4. We leave the treatment on relations as future work.

3.3 Embedding Layer

We use a word embedding layer to convert each token t in X^i into a vector representation $\mathbf{t} \in \mathbb{R}^d$, where d denotes the size of word embedding. To encode positional information, the position encoding (Vaswani et al., 2017) is added as follows:

$$\mathbf{t} = Embed(t) + Pos(t). \quad (2)$$

Similarly, we use a concept embedding layer to convert each concept c into a vector representation $\mathbf{c} \in \mathbb{R}^d$ but without position encoding.

3.4 Dynamic Context-Aware Affective Graph Attention

To enrich word embedding with concept representations, we propose a dynamic context-aware affective graph attention mechanism to compute the concept representation for each token. Specifically, the concept representation $\mathbf{c}(t) \in \mathbb{R}^d$ for token t is computed as

$$\mathbf{c}(t) = \sum_{k=1}^{|g(t)|} \alpha_k * \mathbf{c_k}, \quad (3)$$

where $\mathbf{c_k} \in \mathbb{R}^d$ denotes the concept embedding of c_k and α_k denotes its attention weight. If $|g(t)| = 0$, we set $\mathbf{c}$(t) to the average of all concept embeddings. The attention α_k in Equation 3 is computed as

$$\alpha_k = softmax(w_k), \quad (4)$$

where w_k denotes the weight of c_k.

The derivation of w_k is crucial because it regulates the contribution of $\mathbf{c}_k$ towards enriching $\mathbf{t}$. A standard graph attention mechanism (Velikovi et al., 2018) computes w_k by feeding $\mathbf{t}$ and $\mathbf{c}_k$ into a single-layer feedforward neural network. However, not all related concepts are equal in detecting emotions given the conversational context. In our model, we make the assumption that important concepts are those that relate to the conversational context and have strong emotion intensity. To this end, we propose a context-aware affective graph attention mechanism by incorporating two factors when computing w_k, namely relatedness and affectiveness.

Relatedness: Relatedness measures the strength of the relation between c_k and the conversational context. The relatedness factor in w_k is computed as

$$rel_k = \textit{min-max}(s_k) * \textit{abs}(\cos(\mathbf{CR}(X^i), \mathbf{c_k})), \quad (5)$$

where s_k is the confidence score introduced in Section 3.2, *min-max* denotes min-max scaling for each token t, *abs* denotes the absolute function, cos denotes the cosine similarity function, and $\mathbf{CR}(X^i) \in \mathbb{R}^d$ denotes the context representation of the ith conversation X^i. Here we compute $\mathbf{CR}(X^i)$ as the average of all sentence representations in X^i as follows:

$$\mathbf{CR}(X^i) = \textit{avg}(\mathbf{SR}(X^i_{j-M}), ..., \mathbf{SR}(X^i_j)), \quad (6)$$

where $\mathbf{SR}(X^i_j) \in \mathbb{R}^d$ denotes the sentence representation of X^i_j. We compute $\mathbf{SR}(X^i_j)$ via hierarchical pooling (Shen et al., 2018) where n-gram ($n \leq 3$) representations in X^i_j are first computed by max-pooling and then all n-gram representations are averaged. The hierarchical pooling mechanism preserves word order information to certain degree and has demonstrated superior performance than average pooling or max-pooling on sentiment analysis tasks (Shen et al., 2018).

Affectiveness: Affectiveness measures the emotion intensity of c_k. The affectiveness factor in w_k is computed as

$$\textit{aff}_k = \textit{min-max}(||[V(c_k) - 1/2, A(c_k)/2]||_2), \quad (7)$$

where $||.||_k$ denotes l_k norm, $V(c_k) \in [0, 1]$ and $A(c_k) \in [0, 1]$ denote the valence and arousal values of $VAD(c_k)$, respectively. Intuitively, $\textit{aff}_k$ considers the deviations of valence from neutral and the level of arousal from calm. There is no established method in the literature to compute the emotion intensity based on VAD values, but empirically we found that our method correlates better with an emotion intensity lexicon comprising 6K English words (Mohammad, 2018b) than other methods such as taking dominance into consideration or taking l_1 norm. For concept c_k not in NRC_VAD, we set $\textit{aff}_k$ to the mid value of 0.5.

Combining both rel_k and $\textit{aff}_k$, we define the weight w_k as follows:

$$w_k = \lambda_k * rel_k + (1 - \lambda_k) * \textit{aff}_k, \quad (8)$$

where λ_k is a model parameter balancing the impacts of relatedness and affectiveness on computing concept representations. Parameter λ_k can be fixed or learned during training. The analysis of λ_k is discussed in Section 5.2.

Finally, the concept-enriched word representation $\hat{\mathbf{t}}$ can be obtained via a linear transformation:

$$\hat{\mathbf{t}} = \mathbf{W}[\mathbf{t}; \mathbf{c}(t)], \quad (9)$$

where $[;]$ denotes concatenation and $\mathbf{W} \in \mathbb{R}^{d \times 2d}$ denotes a model parameter. All m tokens in each X^i_j then form a concept-enriched utterance embedding $\hat{\mathbf{X}}^i_j \in \mathbb{R}^{m \times d}$.

3.5 Hierarchical Self-Attention

We propose a hierarchical self-attention mechanism to exploit the structural representation of conversations and learn a vector representation for the contextual utterances $X^i_{j-1}, ..., X^i_{j-M}$. Specifically, the hierarchical self-attention follows two steps: 1) each utterance representation is computed using an utterance-level self-attention layer, and 2) a context representation is computed from M learned utterance representations using a context-level self-attention layer.

At step 1, for each utterance X^i_n, n=$j-1$, ..., $j-M$, its representation $\hat{\mathbf{X}}'^i_n \in \mathbb{R}^{m \times d}$ is learned as follows:

$$\hat{\mathbf{X}}'^i_n = \textit{FF}(L'(\textit{MH}(L(\hat{\mathbf{X}}^i_n), L(\hat{\mathbf{X}}^i_n), L(\hat{\mathbf{X}}^i_n)))), \quad (10)$$

where $L(\hat{\mathbf{X}}^i_n) \in \mathbb{R}^{m \times h \times d_s}$ is linearly transformed from $\hat{\mathbf{X}}^i_n$ to form h heads ($d_s = d/h$), L' linearly transforms from h heads back to 1 head, and

$$\textit{MH}(Q, K, V) = \textit{softmax}(\frac{QK^T}{\sqrt{d_s}})V, \quad (11)$$

$$\textit{FF}(x) = \textit{max}(0, xW_1 + b_1)W_2 + b_2, \quad (12)$$

where Q, K, and V denote sets of queries, keys and values, respectively, $W_1 \in \mathbb{R}^{d \times p}, b_1 \in \mathbb{R}^p, W_2 \in \mathbb{R}^{p \times d}$ and $b_2 \in \mathbb{R}^d$ denote model parameters, and p denotes the hidden size of the point-wise feedforward layer (FF) (Vaswani et al., 2017). The multi-head self-attention layer (MH) enables our model to jointly attend to information from different representation subspaces (Vaswani et al., 2017). The scaling factor $\frac{1}{\sqrt{d_s}}$ is added to ensure the dot product of two vectors do not get overly large. Similar to (Vaswani et al., 2017), both MH and FF layers are followed by residual connection and layer normalization, which are omitted in Equation 10 for brevity.

Dataset	Domain	#Conv. (Train/Val/Test)	#Utter. (Train/Val/Test)	#Classes	Evaluation
EC	Tweet	30160/2755/5509	90480/8265/16527	4	Micro-F1
DailyDialog	Daily Communication	11118/1000/1000	87170/8069/7740	7	Micro-F1
MELD	TV Show Scripts	1038/114/280	9989/1109/2610	7	Weighted-F1
EmoryNLP	TV Show Scripts	659/89/79	7551/954/984	7	Weighted-F1
IEMOCAP	Emotional Dialogues	100/20/31	4810/1000/1523	6	Weighted-F1

Table 1: Dataset descriptions.

At step 2, to effectively combine all utterance representations in the context, the context-level self-attention layer is proposed to hierarchically learn the context-level representation $\mathbf{C}^i \in \mathbb{R}^{M\times m\times d}$ as follows:

$$\mathbf{C}^i = FF(L^{'}(MH(L(\hat{\mathbf{X}}^i), L(\hat{\mathbf{X}}^i), L(\hat{\mathbf{X}}^i)))), \quad (13)$$

where $\hat{\mathbf{X}}^i$ denotes $[\hat{\mathbf{X}}^{'i}_{j-M}; ...; \hat{\mathbf{X}}^{'i}_{j-1}]$, which is the concatenation of all learned utterance representations in the context.

3.6 Context-Response Cross-Attention

Finally, a context-aware concept-enriched response representation $\mathbf{R}^i \in \mathbb{R}^{m\times d}$ for conversation X^i is learned by cross-attention (Bahdanau et al., 2014), which selectively attends to the concept-enriched context representation as follows:

$$\mathbf{R}^i = FF(L^{'}(MH(L(\hat{\mathbf{X}}^{'i}_j), L(\mathbf{C}^i), L(\mathbf{C}^i)))), \quad (14)$$

where the response utterance representation $\hat{\mathbf{X}}^{'i}_j \in \mathbb{R}^{m\times d}$ is obtained via the MH layer:

$$\hat{\mathbf{X}}^{'i}_j = L^{'}(MH(L(\hat{\mathbf{X}}^i_j), L(\hat{\mathbf{X}}^i_j), L(\hat{\mathbf{X}}^i_j))), \quad (15)$$

The resulted representation $\mathbf{R}^i \in \mathbb{R}^{m\times d}$ is then fed into a max-pooling layer to learn discriminative features among the positions in the response and derive the final representation $\mathbf{O} \in \mathbb{R}^d$:

$$\mathbf{O} = max\_pool(\mathbf{R}^i). \quad (16)$$

The output probability p is then computed as

$$p = softmax(\mathbf{O}W_3 + b_3), \quad (17)$$

where $W_3 \in \mathbb{R}^{d\times q}$ and $b_3 \in \mathbb{R}^q$ denote model parameters, and q denotes the number of classes. The entire KET model is optimized in an end-to-end manner as defined in Equation 1. Our model is available at here[1].

[1]https://github.com/zhongpeixiang/KET

4 Experimental Settings

In this section we present the datasets, evaluation metrics, baselines, our model variants, and other experimental settings.

4.1 Datasets and Evaluations

We evaluate our model on the following five emotion detection datasets of various sizes and domains. The statistics are reported in Table 1.

EC (Chatterjee et al., 2019): Three-turn Tweets. The emotion labels include happiness, sadness, anger and other.

DailyDialog (Li et al., 2017): Human written daily communications. The emotion labels include neutral and Ekman's six basic emotions (Ekman, 1992), namely happiness, surprise, sadness, anger, disgust and fear.

MELD (Poria et al., 2018): TV show scripts collected from *Friends*. The emotion labels are the same as the ones used in DailyDialog.

EmoryNLP (Zahiri and Choi, 2018): TV show scripts collected from *Friends* as well. However, its size and annotations are different from MELD. The emotion labels include neutral, sad, mad, scared, powerful, peaceful, and joyful.

IEMOCAP (Busso et al., 2008): Emotional dialogues. The emotion labels include neutral, happiness, sadness, anger, frustrated, and excited.

In terms of the evaluation metric, for EC and DailyDialog, we follow (Chatterjee et al., 2019) to use the micro-averaged F1 excluding the majority class (neutral), due to their extremely unbalanced labels (the percentage of the majority class in the test set is over 80%). For the rest relatively balanced datasets, we follow (Majumder et al., 2019) to use the weighted macro-F1.

4.2 Baselines and Model Variants

For a comprehensive performance evaluation, we compare our model with the following baselines:

cLSTM: A contextual LSTM model. An utterance-level bidirectional LSTM is used to encode each utterance. A context-level unidirectional LSTM is used to encode the context.

Model	EC	DailyDialog	MELD	EmoryNLP	IEMOCAP
cLSTM	0.6913	0.4990	0.4972	0.2601	0.3484
CNN (Kim, 2014)	0.7056	0.4934	0.5586	0.3259	0.5218
CNN+cLSTM (Poria et al., 2017)	0.7262	0.5024	0.5687	0.3289	0.5587
BERT_BASE (Devlin et al., 2018)	0.6946	0.5312	0.5621	0.3315	0.6119
DialogueRNN (Majumder et al., 2019)	0.7405	0.5065	0.5627	0.3170	**0.6121**
KET_SingleSelfAttn (ours)	0.7285	0.5192	0.5624	0.3251	0.5810
KET_StdAttn (ours)	**0.7413**	0.5254	0.5682	0.3353	0.5861
KET (ours)	0.7348	**0.5337**	**0.5818**	**0.3439**	0.5956

Table 2: Performance comparisons on the five test sets. Best values are highlighted in bold.

Dataset	M	m	d	p	h
EC	2	30	200	100	4
DailyDialog	6	30	300	400	4
MELD	6	30	200	100	4
EmoryNLP	6	30	100	200	4
IEMOCAP	6	30	300	400	4

Table 3: Hyper-parameter settings for KET. M: context length. m: number of tokens per utterance. d: word embedding size. p: hidden size in FF layer. h: number of heads.

CNN (Kim, 2014): A single-layer CNN with strong empirical performance. This model is trained on the utterance-level without context.

CNN+cLSTM (Poria et al., 2017): An CNN is used to extract utterance features. An cLSTM is then applied to learn context representations.

BERT_BASE (Devlin et al., 2018): Base version of the state-of-the-art model for sentiment classification. We treat each utterance with its context as a single document. We limit the document length to the last 100 tokens to allow larger batch size. We do not experiment with the large version of BERT due to memory constraint of our GPU.

DialogueRNN (Majumder et al., 2019): The state-of-the-art model for emotion detection in textual conversations. It models both context and speakers information. The CNN features used in DialogueRNN are extracted from the carefully tuned CNN model. For datasets without speaker information, i.e., EC and DailyDialog, we use two speakers only. For MELD and EmoryNLP, which have 260 and 255 speakers, respectively, we additionally experimented with clipping the number of speakers to the most frequent ones (6 main speakers + an universal speaker representing all other speakers) and reported the best results.

KET_SingleSelfAttn: We replace the hierarchical self-attention by a single self-attention layer to learn context representations. Contextual utterances are concatenated together prior to the single self-attention layer.

KET_StdAttn: We replace the dynamic context-aware affective graph attention by the standard graph attention (Velikovi et al., 2018).

4.3 Other Experimental Settings

We preprocessed all datasets by lower-casing and tokenization using Spacy[2]. We keep all tokens in the vocabulary[3]. We use the released code for BERT_BASE and DialogueRNN. For each dataset, all models are fine-tuned based on their performance on the validation set.

For our model in all datasets, we use Adam optimization (Kingma and Ba, 2014) with a batch size of 64 and learning rate of 0.0001 throughout the training process. We use GloVe embedding (Pennington et al., 2014) for initialization in the word and concept embedding layers[4]. For the class weights in cross-entropy loss for each dataset, we set them as the ratio of the class distribution in the validation set to the class distribution in the training set. Thus, we can alleviate the problem of unbalanced dataset. The detailed hyper-parameter settings for KET are presented in Table 3.

5 Result Analysis

In this section we present model evaluation results, model analysis, and error analysis.

5.1 Comparison with Baselines

We compare the performance of KET against that of the baseline models on the five afore-introduced datasets. The results are reported in Table 2. Note that our results for CNN, CNN+cLSTM and DialogueRNN on EC, MELD and IEMOCAP are slightly different from the reported results in (Majumder et al., 2019; Poria et al., 2019).

[2]https://spacy.io/

[3]We keep tokens with minimum frequency of 2 for DailyDialog due to its large vocabulary size

[4]We use GloVe embeddings from Magnitude Medium: https://github.com/plasticityai/magnitude

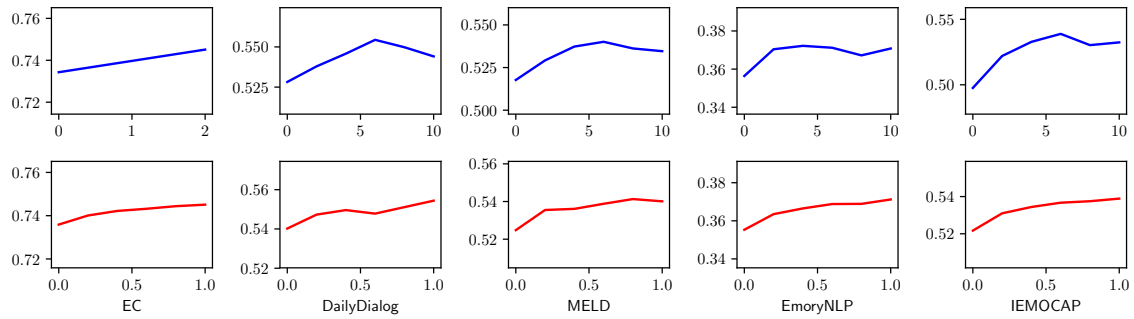

Figure 3: Validation performance by KET. Top: different context length (M). Bottom: different sizes of random fractions of ConceptNet.

cLSTM performs reasonably well on short conversations (i.e., EC and DailyDialog), but the worst on long conversations (i.e., MELD, EmoryNLP and IEMOCAP). One major reason is that learning long dependencies using gated RNNs may not be effective enough because the gradients are expected to propagate back through inevitably a huge number of utterances and tokens in sequence, which easily leads to the vanishing gradient problem (Bengio et al., 1994). In contrast, when the utterance-level LSTM in cLSTM is replaced by features extracted by CNN, i.e., the CNN+cLSTM, the model performs significantly better than cLSTM on long conversations, which further validates that modelling long conversations using only RNN models may not be sufficient. BERT_BASE achieves very competitive performance on all datasets except EC due to its strong representational power via bi-directional context modelling using the Transformer. Note that BERT_BASE has considerably more parameters than other baselines and our model (110M for BERT_BASE versus 4M for our model), which can be a disadvantage when deployed to devices with limited computing power and memory. The state-of-the-art DialogueRNN model performs the best overall among all baselines. In particular, DialogueRNN performs better than our model on IEMOCAP, which may be attributed to its detailed speaker information for modelling the emotion dynamics in each speaker as the conversation flows.

It is encouraging to see that our KET model outperforms the baselines on most of the datasets tested. This finding indicates that our model is robust across datasets with varying training sizes, context lengths and domains. Our KET variants KET_SingleSelfAttn and KET_StdAttn perform comparably with the best baselines on all datasets except IEMOCAP. However, both variants perform noticeably worse than KET on all datasets except EC, validating the importance of our proposed hierarchical self-attention and dynamic context-aware affective graph attention mechanism. One observation worth mentioning is that these two variants perform on a par with the KET model on EC. Possible explanations are that 1) hierarchical self-attention may not be critical for modelling short conversations in EC, and 2) the informal linguistic styles of Tweets in EC, e.g., misspelled words and slangs, hinder the context representation learning in our graph attention mechanism.

5.2 Model Analysis

We analyze the impact of different settings on the validation performance of KET. All results in this section are averaged over 5 random seeds.

Analysis of context length: We vary the context length M and plot model performance in Figure 3 (top portion). Note that EC has only a maximum number of 2 contextual utterances. It is clear that incorporating context into KET improves performance on all datasets. However, adding more context is contributing diminishing performance gain or even making negative impact in some datasets. This phenomenon has been observed in a prior study (Su et al., 2018). One possible explanation is that incorporating long contextual information may introduce additional noises, e.g., polysemes expressing different meanings in different utterances of the same context. More thorough investigation of this diminishing return phenomenon is a worthwhile direction in the future.

Analysis of the size of ConceptNet: We vary the size of ConceptNet by randomly keeping only a fraction of the concepts in ConceptNet when train-

Dataset	0	0.3	0.7	1
EC	0.7345	0.7397	**0.7426**	0.7363
DailyDialog	0.5365	0.5432	**0.5451**	0.5383
MELD	0.5321	**0.5395**	0.5366	0.5306
EmoryNLP	0.3528	**0.3624**	0.3571	0.3488
IEMOCAP	0.5344	**0.5367**	0.5314	0.5251

Table 4: Analysis of the relatedness-affectiveness tradeoff on the validation sets. Each column corresponds to a fixed λ_k for all concepts (see Equation 8).

Dataset	KET	-context	-knowledge
EC	**0.7451**	0.7343	0.7359
DailyDialog	**0.5544**	0.5282	0.5402
MELD	**0.5401**	0.5177	0.5248
EmoryNLP	**0.3712**	0.3564	0.3553
IEMOCAP	**0.5389**	0.4976	0.5217

Table 5: Ablation study for KET on the validation sets.

ing and evaluating our model. The results are illustrated in Figure 3 (bottom portion). Adding more concepts consistently improves model performance before reaching a plateau, validating the importance of commonsense knowledge in detecting emotions. We may expect the performance of our KET model to improve with the growing size of ConceptNet in the future.

Analysis of the relatedness-affectiveness tradeoff: We experiment with different values of $\lambda_k \in [0, 1]$ (see Equation 8) for all k and report the results in Table 4. It is clear that λ_k makes a noticeable impact on the model performance. Discarding relatedness or affectiveness completely will cause significant performance drop on all datasets, with one exception of IEMOCAP. One possible reason is that conversations in IEMOCAP are emotional dialogues, therefore, the affectiveness factor in our proposed graph attention mechanism can provide more discriminative power.

Ablation Study: We conduct ablation study to investigate the contribution of context and knowledge as reported in Table 5. It is clear that both context and knowledge are essential to the strong performance of KET on all datasets. Note that removing context has a greater impact on long conversations than short conversations, which is expected because more contextual information is lost in long conversations.

5.3 Error Analysis

Despite the strong performance of our model, it still fails to detect certain emotions on certain datasets. We rank the F1 score of each emotion per dataset and investigate the emotions with the worst scores. We found that disgust and fear are generally difficult to detect and differentiate. For example, the F1 score of fear emotion in MELD is as low as 0.0667. One possible cause is that these two emotions are intrinsically similar. The VAD values of both emotions have low valence, high arousal and low dominance (Mehrabian, 1996). Another cause is the small amount of data available for these two emotions. How to differentiate intrinsically similar emotions and how to effectively detect emotions using limited data are two challenging directions in this field.

6 Conclusion

We present a knowledge-enriched transformer to detect emotions in textual conversations. Our model learns structured conversation representations via hierarchical self-attention and dynamically refers to external, context-aware, and emotion-related knowledge entities from knowledge bases. Experimental analysis demonstrates that both contextual information and commonsense knowledge are beneficial to model performance. The tradeoff between relatedness and affectiveness plays an important role as well. In addition, our model outperforms the state-of-the-art models on most of the tested datasets of varying sizes and domains.

Given that there are similar emotion lexicons to NRC_VAD in other languages and ConceptNet is a multilingual knowledge base, our model can be easily adapted to other languages. In addition, given that NRC_VAD is the only emotion-specific component, our model can be adapted as a generic model for conversation analysis.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments. This research is supported, in part, by the National Research Foundation, Prime Ministers Office, Singapore under its AI Singapore Programme (Award Number: AISG-GC-2019-003) and under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). This research is also supported, in part, by the Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *ACL*, volume 1, pages 718–728.

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *ECIR*, pages 154–166. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak Padmanabhan. 2017. Lexicon generation for emotion detection from text. *IEEE Intelligent Systems*, 32(1):102–108.

Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

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

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

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *EMNLP*, pages 551–561.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Laurence Devillers, Ioana Vasilescu, and Lori Lamel. 2002. Annotation and detection of emotion in a task-oriented human-human dialog corpus. In *Proceedings of ISLE Workshop*.

Laurence Devillers and Laurence Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth International Conference on Spoken Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

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

Christiane Fellbaum. 2012. Wordnet. *The Encyclopedia of Applied Linguistics*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th SIGDIAL*, pages 129–133.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *ACL*, pages 221–231.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *EMNLP*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *NAACL*, volume 1, pages 2122–2132.

Junqing He, Bing Wang, Mingming Fu, Tianqi Yang, and Xuemin Zhao. 2019. Hierarchical attention and knowledge matching networks with information enhancement for end-to-end task-oriented dialog systems. *IEEE Access*, 7:18871–18883.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yajie Hu, Xiaoou Chen, and Deshun Yang. 2009. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*, pages 123–128.

Eva Hudlicka. 2011. Guidelines for designing computational models of emotions. *International Journal of Synthetic Emotions*, 2(1):26–79.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *EMNLP*, pages 1160–1166.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *EMNLP*, pages 329–339.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Abhishek Kumar, Daisuke Kawahara, and Sadao Kurohashi. 2018. Knowledge-enriched two-layered attention network for sentiment analysis. In *NAACL*, volume 2, pages 253–258.

Chul Min Lee and Shrikanth S Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, volume 1, pages 986–995.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *ACL*, pages 1489–1498.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *ACL*, volume 1, pages 1468–1478.

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

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *ACL*, pages 821–832.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *EMNLP*, pages 2322–2332.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *ACL*, pages 174–184.

Saif M. Mohammad. 2018b. Word affect intensities. In *LREC*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *EMNLP*, pages 690–695.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*, volume 1, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *arXiv preprint arXiv:1905.02947*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Yi Luan Mirella Lapata Rik Koncel-Kedziorski, Dhanush Bekal and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. In *NAACL*.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *ACL*, pages 440–450.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen. 2018. How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues. In *NAACL*, volume 1, pages 2133–2142.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *EMNLP*, pages 4231–4242.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

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

Petar Velikovi, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL*, pages 90–94. Association for Computational Linguistics.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *ICLR*.

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

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at AAAI*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018a. Improving the transformer translation model with document-level context. In *EMNLP*, pages 533–542.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018b. Text emotion distribution learning via multi-task convolutional neural network. In *IJCAI*, pages 4595–4601.

Peixiang Zhong and Chunyan Miao. 2019. ntuer at SemEval-2019 task 3: Emotion classification with word and sentence representations in RCNN. In *SemEval*, pages 282–286.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *AAAI*, pages 7492–7500.

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

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018c. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*, volume 1, pages 1118–1127.