Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection

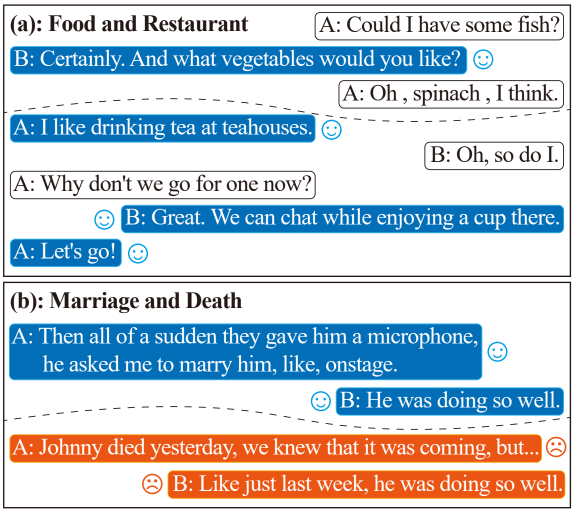
Lixing Zhu[†], Gabriele Pergola[†], Lin Gui[†], Deyu Zhou[§], Yulan He[†]
[†]Department of Computer Science, University of Warwick, UK
[§]School of Computer Science and Engineering, Key Laboratory of Computer Network
and Information Integration, Ministry of Education, Southeast University, China
{lixing.zhu,gabriele.pergola,lin.gui,yulan.he}@warwick.ac.uk
d.zhou@seu.edu.cn

Abstract

Emotion detection in dialogues is challenging as it often requires the identification of thematic topics underlying a conversation, the relevant commonsense knowledge, and the intricate transition patterns between the affective states. In this paper, we propose a Topic-Driven Knowledge-Aware Transformer to handle the challenges above. We firstly design a topic-augmented language model (LM) with an additional layer specialized for topic detection. The topic-augmented LM is then combined with commonsense statements derived from a knowledge base based on the dialogue contextual information. Finally, a transformer-based encoder-decoder architecture fuses the topical and commonsense information, and performs the emotion label sequence prediction. The model has been experimented on four datasets in dialogue emotion detection, demonstrating its superiority empirically over the existing state-of-the-art approaches. Quantitative and qualitative results show that the model can discover topics which help in distinguishing emotion categories.

1 Introduction

The abundance in dialogues extracted from online conversations and TV series provides unprecedented opportunity to train models for automatic emotion detection, which are important for the development of empathetic conversational agents or chat bots for psychotherapy (Hsu and Ku, 2018; Jiao et al., 2019; Zhang et al., 2019; Cao et al., 2019). However, it is challenging to capture the contextual semantics of personal experience described in one's utterance. For example, the emotion of the sentence "*I just passed the exam*" can be either *happy* or *sad* depending on the expectation of the subject. There are strands of works utilizing the dialogue context to enhance the utterance representation (Jiao et al., 2019; Zhang et al., 2019; Majumder et al., 2019), where influences from historical utterances were handled by recurrent units, and attention signals were further introduced to intensify the positional order of the utterances.


(a): Food and Restaurant

A: Could I have some fish?

B: Certainly. And what vegetables would you like? ☺

A: Oh , spinach , I think.

A: I like drinking tea at teahouses. ☺

B: Oh, so do I.

A: Why don't we go for one now?

☺ B: Great. We can chat while enjoying a cup there.

A: Let's go! ☺

(b): Marriage and Death

A: Then all of a sudden they gave him a microphone, he asked me to marry him, like, onstage. ☺

☺ B: He was doing so well.

A: Johnny died yesterday, we knew that it was coming, but... ☹

☹ B: Like just last week, he was doing so well.

Figure 1: Utterances around particular topics carry specific emotions. Utterances carrying *positive* (smiling face) or *negative* (crying face) emotions are highlighted in colour. Other utterances are labeled as '*Neutral*'. In (a), utterances discussing food and restaurant are more likely carrying positive sentiment. In (b), the similar utterance, '*He was doing so well*', expressed different emotions depending on its associated topic.

Despite the progress made by the aforementioned methods, detecting emotions in dialogues is however still a challenging task due to the way emotions are expressed and how the meanings of utterances vary based on the particular topic discussed, as well as the implicit knowledge shared between participants. Figure 1 gives an example of how topics and background knowledge could impact the mood of interlocutors. Normally, dialogues around specific topics carry certain language patterns (Serban et al., 2017), affecting not only the utterance's meaning, but also the particular emo-

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 1571–1582
August 1–6, 2021. ©2021 Association for Computational Linguistics

tions conveyed by specific expressions. Existing dialogue emotion detection methods did not put emphasis on modelling these holistic properties of dialogues (i.e., conversational topics and tones). Consequently, they were fundamentally limited in capturing the affective states of interlocutors related to the particular themes discussed. Besides, emotion and topic detection heavily relies on leveraging underlying commonsense knowledge shared between interlocutors. Although there have been attempts in incorporating it, such as the COSMIC (Ghosal et al., 2020), existing approaches do not perform fine-grained extraction of relevant information based on both the topics and the emotions involved.

Recently, the Transformer architecture (Vaswani et al., 2017) has empowered language models to transfer large quantities of data to low-resource domains, making it viable to discover topics in conversational texts. In this paper, we propose to add an extra layer to the pre-trained language model to model the latent topics, which is learned by fine-tuning on dialogue datasets to alleviate the data sparsity problem. Inspired by the success of Transformers, we use the Transformer Encoder-Decoder structure to perform the Seq2Seq prediction in which an emotion label sequence is predicted given an utterance sequence (i.e., each utterance is assigned with an emotion label). We posit that the dialogue emotion of the current utterance depends on the historical dialogue context and the predicted emotion label sequence for the past utterances. We leverage the attention mechanism and the gating mechanism to incorporate commonsense knowledge retrieved by different approaches. Code and trained models are released to facilitate further research[1]. To sum up, our contributions are:

- We are the first to propose a topic-driven approach for dialogue emotion detection. We propose to alleviate the low-resource setting by topic-driven fine-tuning using pre-trained language models.
- We utilize a pointer network and an additive attention to integrate commonsense knowledge from multiple sources and dimensions.
- We develop a Transformer Encoder-Decoder structure as a replacement of the commonly-used recurrent attention neural networks for dialogue emotion detection.

2 Related Work

Dialogue Emotion Detection Majumder et al. (2019) recognized the importance of dialogue context in dialogue emotion detection. They used a Gated Recurrent Unit (GRU) to capture the global context which is updated by the speaker ad-hoc GRUs. At the same time, Jiao et al. (2019) presented a hierarchical neural network model that comprises two GRUs for the modelling of tokens and utterances respectively. Zhang et al. (2019) explicitly modelled the emotional dependencies on context and speakers using a Graph Convolutional Network (GCN). Meanwhile, Ghosal et al. (2019) extended the prior work (Majumder et al., 2019) by taking into account the intra-speaker dependency and relative position of the target and context within dialogues. Memory networks have been explored in (Jiao et al., 2020) to allow bidirectional influence between utterances. A similar idea has been explored by Li et al. (2020b). While the majority of works have been focusing on textual conversations, Zhong et al. (2019) enriched utterances with concept representations extracted from the ConceptNet (Speer et al., 2017). Ghosal et al. (2020) developed COSMIC which exploited ATOMIC (Sap et al., 2019) for the acquisition of commonsense knowledge. Different from existing approaches, we propose a topic-driven and knowledge-aware model built on a Transformer Encoder-Decoder structure for dialogue emotion detection.

Latent Variable Models for Dialogue Context Modelling Latent variable models, normally described in their neural variational inference form named Variational Autoencoder (VAE) (Kingma and Welling, 2014), has been studied extensively to learn thematic representations of individual documents (Miao et al., 2016; Srivastava and Sutton, 2017; Rezaee and Ferraro, 2020). They have been successfully employed for dialogue generation to model thematic characteristics over dynamically evolving conversations. This line of work, which inlcudes approaches based on hierarchical recurrent VAEs (Serban et al., 2017; Park et al., 2018; Zeng et al., 2019) and conditional VAEs (Sohn et al., 2015; Shen et al., 2018; Gao et al., 2019), encodes each utterance with historical latent codes and autoregressively reconstructs the input sequence.

On the other hand, pre-trained language models are used as embedding inputs to VAE-based mod-

[1] http://github.com/something678/TodKat.

els (Peinelt et al., 2020; Asgari-Chenaghlu et al., 2020). Recent work by Li et al. (2020a) employs BERT and GPT-2 as the encoder-decoder structure of VAE. However, these models have to be either trained from scratch or built upon pre-trained embeddings. They therefore cannot be directly applied to the low-resource setting of dialogue emotion detection.

Knowledge Base and Knowledge Retrieval ConceptNet (Speer et al., 2017) captures commonsense concepts and relations as a semantic network, which encompasses the spatial, physical, social, temporal, and psychological aspects of everyday life. More recently, Sap et al. (2019) built ATOMIC, a knowledge graph centered on events rather than entities. Owing to the expressiveness of events and ameliorated relation types, using ATOMIC achieved competitive results against human evaluation in the task of If-Then reasoning.

Alongside the development of knowledge bases, recent years have witnessed the thrive of new methods for training language models from large-scale text corpora as implicit knowledge base. As has been shown in (Petroni et al., 2019), pre-trained language models perform well in recalling relational knowledge involving triplet relations about entities. Bosselut et al. (2019) proposed COMmonsEnse Transformers (COMET) which learns to generate commonsense descriptions in natural language by fine-tuning pre-trained language models on existing commonsense knowledge bases such as ATOMIC. Compared with extractive methods, language models fine-tuned on knowledge bases have a distinctive advantage of being able to generate knowledge for unseen events, which is of great importance for tasks which require the incorporation of commonsense knowledge such as emotion detection in dialogues.

3 Methodology

3.1 Problem Setup

A dialogue is defined as a sequence of utterances $\{x_1, x_2, \ldots, x_N\}$, which is annotated with a sequence of emotion labels $\{y_1, y_2, \ldots, y_N\}$. Our goal is to develop a model that can assign the correct label to each utterance. As for each utterance, the raw input is a token sequence, i.e., $x_n = \{w_{n,1}, w_{n,2}, \ldots, w_{n,M_n}\}$ where M_n denotes the length of an utterance. We address this problem using the Seq2Seq framework (Sutskever et al., 2014), in which the model consecutively consumes an utterance x_n and predicts the emotion label y_n based on the earlier utterances and their associated predicted emotion labels. The joint probability of emotion labels for a dialogue is:

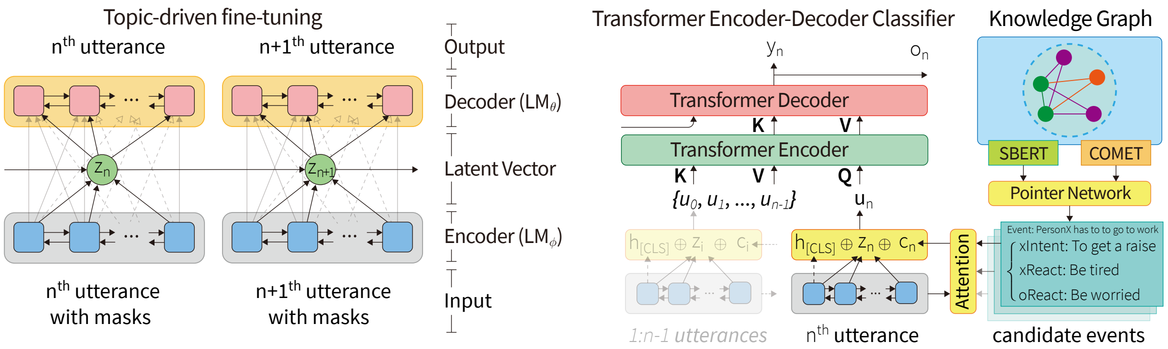
$$P_\theta(y_{1:N}|x_{1:N}) = \prod_{n=1}^{N} P_\theta(y_n|x_{\leq n}, y_{<n}) \quad (1)$$

It is worth mentioning that the subsequent utterances are unseen to the model at each predictive step. Learning is performed via optimizing the log-likelihoods of predicted emotion labels.

The overall architecture of our proposed TOpic-Driven and Knowledge-Aware Transformer (TODKAT) is shown in Figure 2, which consists of two main components, the topic-driven language model fine tuned on dialogues, and the knowledge-aware transformer for emotion label sequence prediction for a given dialogue. In what follows, we will describe each of the components in turn.

3.2 Topic Representation Learning

We propose to insert a topic layer into an existing language model and fine-tune the pre-trained language model on the conversational text for topic representation learning. Topic models, often formulated as latent variable models, play a vital role in dialogue modeling (Serban et al., 2017) due to the explicit modeling of 'high-level syntactic features such as style and topic' (Bowman et al., 2016). Despite the tremendous success of applying topic modeling in dialogue generation (Sohn et al., 2015; Shen et al., 2018; Gao et al., 2019), there is scarce work exploiting latent variable models for dialogue emotion detection. To this end, we borrow the architecture from VHRED (Serban et al., 2017) for topic discovery, with the key modification that both the encoder RNN and decoder RNN are replaced by layers of a pre-trained language model. Furthermore, we use a transformer multi-head attention in replacement of the LSTM to model the dependence between the latent topic vectors. Unlike VHRED, we are interested in the encoder part to extract the posterior of the latent topic z, rather than the recurrent prior of z in the decoder part since the latter is intended for dialogue generation. We assume that each utterance is mapped to a latent variable encoding its internal topic, and impose a sequential dependence on the topic transitions. Figure 2a gives an overview of the VAE-based model which



(a) Topic-driven fine-tuning of a pre-trained LM.

(b) Knowledge-aware transformer.

Figure 2: TOpic-Driven and Knowledge-Aware Transformer (TODKAT).

aims at learning the latent topic vector during the fine-tuning of the language model.

Specifically, the pre-trained language model is decomposed into two parts, the encoder and the decoder. By retaining the pre-trained weights, we transfer representations from high-resource tasks to the low-resource setting, which is the case for dialogue emotion datasets.

Encoder The training of topic discovery part of TODKAT comprises a VAE at each time step, with its latent variable dependent on the previous latent code. Each utterance is input to the VAE encoder with a recurrent hidden state, the output of which is a latent vector ideally encoding the topic discussed in the utterance. The latent vectors are tied through a recurrent hidden state to constraint a coherent topic over a single dialogue. We use LM_ϕ to denote the network of lower layers of the language model (before the topic layer) and x_n^L to denote the output from LM_ϕ given the input x_n. The variational distribution for the approximation of the posterior will be:

$$\begin{aligned} &q_\phi(z_n|\boldsymbol{x}_{\le n}, \boldsymbol{z}_{<n}) \\ &= \mathcal{N}\big(z_n|f_{\mu_\phi}(x_n^L, h_{n-1}), f_{\sigma_\phi}(x_n^L, h_{n-1})\big), \end{aligned} \quad (2)$$

$$\text{where } h_{n-1} = f_\tau(z_{n-1}, x_{n-1}^L), \text{for } n > 1. \quad (3)$$

Here, $f_{\mu_\phi}(\cdot)$ and $f_{\sigma_\phi}(\cdot)$ are multi-layer perceptrons (MLPs), f_τ can be any transition function (e.g., a recurrent unit). We employ the transformer multi-head attention with its query being the previous latent variable z_{n-1}, that is,

$$f_\tau(z_{n-1}, x_{n-1}^L) = \text{Attention}(z_{n-1}, x_{n-1}^L, x_{n-1}^L). \quad (4)$$

We initialize $h_0 = \mathbf{0}$ and model the transition between h_{n-1} and h_n by first generating z_n from h_{n-1} using Eq. (2), then calculating h_n by Eq. (3).

Decoder The decoder network reconstructs x_n from z_n at each time step. We use Gaussian distributions for both the generative prior and the variational distribution. Since we want z_n to be dependent on z_{n-1}, the prior for z_n is $p(z_n|h_{n-1}) = \mathcal{N}\big(z_n|f_{\mu_\gamma}(h_{n-1}), f_{\sigma_\gamma}(h_{n-1})\big)$. where $f_{\mu_\gamma}(\cdot)$ and $f_{\sigma_\gamma}(\cdot)$ are MLPs. The posterior for z_n is $p_\theta(z_n|\boldsymbol{x}_{\le n}, \boldsymbol{z}_{<n})$, which is intractable and is approximated by $q_\phi(z_n|\boldsymbol{x}_{\le n}, \boldsymbol{z}_{<n})$ of Eq. 2. We denote the higher layers of the language model as LM_θ. Then the reconstruction of $\hat{x}_n$ given z_n and x_n^L can be expressed as:

$$\hat{x}_n = \text{LM}_\theta(z_n, x_n^L). \quad (5)$$

Note that this is different from dialogue generation in which an utterance is generated from the latent topic vector. Here, we aim to extract the latent topic from the current utterance and therefore train the model to reconstruct the input utterance as specified in Eq. (5). To make the combination of z_n and x_n^L compatible for LM_θ, we need to perform the latent vector injection. As in (Li et al., 2020a), we employ the "Memory" scheme that z_n becomes an additional input for LM_θ, that is, the input to the higher layers becomes $[z_n, x_n^L]$.

Training The training objective is the Evidence Lower Bound (ELBO):

$$\begin{aligned} &\mathbb{E}_{q_\phi(\boldsymbol{z}_{\le N}|\boldsymbol{x}_{\le N})}[\log p_\theta(\boldsymbol{x}_{\le N}|\boldsymbol{z}_{\le N})] \\ &-\text{KL}[q_\phi(\boldsymbol{z}_{\le N}|\boldsymbol{x}_{\le N})||p(\boldsymbol{z}_{\le N})]. \end{aligned} \quad (6)$$

Eq. 6 factorizes and the expectation term becomes

$$\mathbb{E}_{q_\phi(\boldsymbol{z}_{\le N}|\boldsymbol{x}_{\le N})}\left[\sum_{n=1}^{N} \log p_\theta(x_n|\boldsymbol{z}_{\le n}, \boldsymbol{x}_{<n})\right], \quad (7)$$

and the KL term becomes

$$\sum_{n=1}^{N} \mathrm{KL}[q_\phi(z_n|\boldsymbol{x}_{\leq n}, \boldsymbol{z}_{<n})||p(z_n|\boldsymbol{z}_{<n}, \boldsymbol{x}_{<n})], \quad (8)$$

where $p(z_n|\boldsymbol{z}_{<n}, \boldsymbol{x}_{<n})$ is the prior for z_n. After training, we are able to extract the topic representation from the encoder part of the model, which is denoted as $z_n = \mathrm{LM}_\phi^{\mathrm{enc}}(x_n)$. Meanwhile, the entire language model has been fine-tuned, which is denoted as $u_n = \mathrm{LM}^{\mathrm{CLS}}(x_n)$.

3.3 Knowledge-Aware Transformer

The topic-driven LM fine-tuning stage makes it possible for the LM to discover a topic representation from a given utterance. After fine-tuning, we attach the fine-tuned components to a classifier and train the classifier to predict the emotion labels. We propose to use the Transformer Encoder-Decoder structure as the classifier, and consider the incorporation of commonsense knowledge retrieved from external knowledge sources. In what follows, we first describe how to retrieve the commonsense knowledge from a knowledge source, then we present the detailed structure of the classifier.

Commonsense Knowledge Retrieval We use ATOMIC[2] as a source of external knowledge. In ATOMIC, each node is a phrase describing an event. Edges are relation types linking from one event to another. ATOMIC thus encodes triples such as ⟨`event, relation type, event`⟩. There are a total of nine relation types, of which three are used: `xIntent`, the intention of the subject (e.g., '*to get a raise*'), `xReact`, the reaction of the subject (e.g., '*be tired*'), and `oReact`, the reaction of the object (e.g., '*be worried*'), since they are defined as the mental states of an event (Sap et al., 2019).

Given an utterance x_n, we can compare it with every node in the knowledge graph, and retrieve the most similar one. The method for computing the similarity between an utterance and events is SBERT (Reimers and Gurevych, 2019). We extract the top-K events, and obtain their intentions and reactions, which are denoted as $\{e_{n,k}^{sI}, e_{n,k}^{sR}, e_{n,k}^{oR}\}, k = 1, \ldots, K$.

On the other hand, there is a knowledge generation model, called COMET[3], which is trained on ATOMIC. It can take x_n as input and generate the knowledge with the desired event relation types specified (e.g., `xIntent`, `xReact` or `oReact`). The generated knowledge can be unseen in ATOMIC since COMET is essentially a fine-tuned language model. We use COMET to generate the K most likely events, each with respect to the three event relation types. The produced events are denoted as $\{g_{n,k}^{sI}, g_{n,k}^{sR}, g_{n,k}^{oR}\}$, $k = 1, \ldots, K$.

Knowledge Selection With the knowledge retrieved from ATOMIC, we build a pointer network (Vinyals et al., 2015) to exclusively choose the commonsense knowledge either from SBERT or COMET. The pointer network calculates the probability of choosing the candidate knowledge source as:

$$P\big(\mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n) = 1\big) = \sigma\big([x_n, \boldsymbol{e}_n, \boldsymbol{g}_n]\mathbf{W}_\sigma\big),$$

where $\mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n)$ is an indicator function with value 1 or 0, and $\sigma(x) = 1/(1+\exp(-x))$. We envelope σ with Gumbel Softmax (Jang et al., 2017) to generate the one-hot distribution[4]. The integrated commonsense knowledge is expressed as

$$\boldsymbol{c}_n = \mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n)\boldsymbol{e}_n + \big(1 - \mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n)\big)\boldsymbol{g}_n,$$

where $\boldsymbol{c}_n = \{c_{n,k}^{sI}, c_{n,k}^{sR}, c_{n,k}^{oR}\}_{k=1}^{K}$.

With the knowledge source selected, we proceed to select the most informative knowledge. We design an attention mechanism (Bahdanau et al., 2015) to integrate the candidate knowledge. Recall that we have a fine-tuned language model which can calculate both the `[CLS]` and topic representations. Here we apply the language model to the retrieved or generated knowledge to obtain the `[CLS]` and the topic representation, denoted as $[\boldsymbol{c}_{n,k}, z_{n,k}]$. The attention mechanism is performed by calculating the dot product between the utter-

[2] https://homes.cs.washington.edu/~msap/atomic/

[3] https://github.com/atcbosselut/comet-commonsense

[4] We have also experimented with a soft gating mechanism by aggregating knowledge from SBERT and COMET in a weighted manner. But the results are consistently worse than those using a hard gating mechanism.

ance and each normalized knowledge tuple:

$$v_k = \tanh\big([\boldsymbol{c}_{n,k}, z_{n,k}]\mathbf{W}_\alpha\big), \quad (9)$$

$$\alpha_k = \frac{\exp\big(v_k[z_n, u_n]^\top\big)}{\sum_k \exp\big(v_k[z_n, u_n]^\top\big)}, \quad (10)$$

$$\boldsymbol{c}_n = \sum_{k=1}^{K} \alpha_k \boldsymbol{c}_{n,k}. \quad (11)$$

Here, we abuse $\boldsymbol{c}_n$ to represent the aggregated knowledge phrases. We further aggregate $\boldsymbol{c}_n$ by event relation types using a self-attention and the final event representation is denoted as c_n.

Transformer Encoder-Decoder We use a Transformer encoder-decoder to map an utterance sequence to an emotion label sequence, thus allowing for modeling the transitional patterns between emotions and taking into account the historical utterances as well. Each utterance is converted to the [CLS] representation concatenated with the topic representation z_n and knowledge representation c_n. We enforce a masking scheme in the self-attention layer of the encoder to make the classifier predict emotions in an auto-regressive way, entailing that only the past utterances are visible to the encoder. This masking strategy, preventing the query from attending to future keys, suits better a real-world scenario in which the subsequent utterances are unseen when predicting an emotion of the current utterance. As for the decoder, the output of the previous decoder block is input as a query to the self-attention layer. The training loss for the classifier is the negative log-likelihood expressed as:

$$\mathcal{L} = -\sum_{n=1}^{N} \log p_\theta(y_n | \boldsymbol{u}_{\leq n}, \boldsymbol{y}_{<n}),$$

where θ denotes the trainable parameters.

4 Experimental Setup

In this section, we present the details of the datasets used, the methods for comparison, and the implementation details of our models.

Datasets We use the following datasets for experimental evaluation:

DailyDialog (Li et al., 2017) is collected from daily communications. It takes the Ekman's six emotion types (Ekman, 1993) as the annotation protocol, that is, it annotates an utterance with one of the six basic emotions: *anger, disgust, fear, happiness, sadness*, or *surprise*. Those showing ambiguous emotions are annotated as *neutral*.

MELD (Poria et al., 2019) is constructed from scripts of '*Friends*', a TV series on urban life. Same as DailyDialog, the emotion label falls into Ekman's six emotion types, or *neutral*.

IEMOCAP (Busso et al., 2008) is built with subtitles from improvised videos. Its emotion labels are *happy, sad, neutral, angry, excited* and *frustrated*.

EmoryNLP (Zahiri and Choi, 2018)[5] is also built with conversations from '*Friends*' TV series, but with a slightly different annotation scheme in which *disgust, anger* and *surprise* become *peaceful, mad* and *powerful*, respectively.

Following Zhong et al. (2019) and Ghosal et al. (2020), the '*neutral*' label of DailyDialog is not counted in the evaluation to avoid highly imbalanced classes. For MELD and EmoryNLP, we consider a dialogue as a sequence of utterances from the same scene ID. Table 1 summarizes the statistics of each dataset.

	DD	MELD	IEMOCAP	EmoryNLP
#Dial.	13,118	1,432	151	827
Train	11,118	1,038	100	659
Dev.	1,000	114	20	89
Test	1,000	280	31	79
#Utt.	102,979	13,708	7,333	9,489
Train	87,170	9,989	4,810	7,551
Dev.	8,069	1,109	1,000	954
Test	7,740	2,610	1,523	984
#Cat.	7	7	6	7

Table 1: Statistics of the benchmarks for dialogue emotion detection. The train/development/test sets are predefined in each dataset.

Baselines We compare the performance of TODKAT with the following methods:

HiGRU (Jiao et al., 2019) simply inherits the recurrent attention framework that an attention layer is placed between two GRUs to aggregate the signals from the encoder GRU and pass them to the decoder GRU.

DialogueGCN (Ghosal et al., 2019) creates a graph from interactions of speakers to take into account the dialogue structure. A Graph Convolutional Network (GCN) is employed to encode the speakers. Emotion labels are predicted with the combinations of the global context and speakers' status.

[5] https://github.com/emorynlp/emotion-detection

Models	DailyDialog		MELD		IEMOCAP		EmoryNLP	
	Macro-F1 - neutral	Micro-F1 - neutral	weighted Avg-F1	Micro-F1	weighted Avg-F1	Micro-F1	weighted Avg-F1	Micro-F1
HiGRU	0.4904	0.5190	0.5681	0.5452	0.5854	0.5828	0.3448	0.3354
DialogueGCN	0.4995	0.5373	0.5837	0.5617	0.6085	0.6063	0.3429	0.3313
KET	–	0.5348	0.5818	–	0.5956	–	0.3439	–
COSMIC	0.5105	**0.5848**	0.6521	–	**0.6528***	–	0.3811	–
TodKat	**0.5256**	0.5847	**0.6823**	**0.6475**	0.6133	0.6111	**0.4312**	**0.4268**
−Topics	0.5136	0.5549	0.6634	0.6352	0.6281	**0.6260**	0.4180	0.4055
−KB	0.5003	0.5344	0.6397	0.6111	0.5896	0.5738	0.3379	0.3262
KAT$_{\text{SBERT}}$	0.5173	0.5578	0.6454	0.6188	0.6097	0.6069	0.3734	0.3567
KAT$_{\text{COMET}}$	0.5102	0.5462	0.6582	0.6307	0.6277	0.6254	0.4110	0.3974

Table 2: The F1 results of the dialogue emotion detectors on four benchmarks. Here we denote the proposed model as TodKat, of which the results are an average of ten runs. The ablations of different components are reported separately in the bottom, where the model without the incorporation of latent topics is denoted as '−Topics', transformer encoder-decoder structure without the use of a knowledge base is dnoted as '−KB'. KAT$_{\text{COMET}}$ and KAT$_{\text{SBERT}}$ uses the commonsense knowledge obtained with Comet and SBERT, respectively. Results of KET and COSMIC are from (Zhong et al., 2019) and (Ghosal et al., 2020), respectively.

KET (Zhong et al., 2019) is the first model which integrates common-sense knowledge extracted from ConceptNet and emotion information from an emotion lexicon into conversational text. A Transformer encoder is employed to handle the influence from past utterances.

COSMIC (Ghosal et al., 2020) is the state-of-the-art approach that leverages Atomic for improved emotion detection. Comet is employed in their model to retrieve the event-eccentric commonsense knowledge from Atomic.

We modified the script[6] of language model fine-tuning in the Hugging Face library (Wolf et al., 2020) for the implementation of topic-driven fine-tuning. We use one transformer encoder layer. As for the decoder, there are N layers where N is the number of utterances in a dialogue. We refer the readers to the Appendix for the detailed settings of the proposed models.

5 Results and Analysis

Comparison with Baselines Experiment results of TodKat and its ablations are reported in Table 2. HiGRU and DialogueGCN results were produced by running the code published by the authors on the four datasets. Among the baselines, COSMIC gives the best results. Our proposed TodKat outperforms COSMIC on both MELD and EmoryNLP in weighted Avg-F1 with the improvements ranging between 3-5%. TodKat also achieves superior result than COSMIC on DailyDialogue in Macro-F1 and gives nearly the same result in Micro-F1. TodKat is inferior to COSMIC on IEMOCAP. It is however worth mentioning that COSMIC was trained with 132 instances on this dataset, while for all the other models the training-and-validation split is 100 and 20. As such, the IEMOCAP results reported on COSMIC (Ghosal et al., 2020) are not directly comparable here. COSMIC also incorporates the commonsense knowledge from Atomic but with the modified GRUs. Our proposed TodKat, built upon the topic-driven Transformer, appears to be a more effective architecure for dialogue emotion detection. Compared with KET, the improvements are much more significant, with over 10% increase on MELD, and close to 5% gain on DailyDialog. KET is also built on the Transformer, but it considers each utterance in isolation and applies commonsense knowledge from ConceptNet. TodKat, on the contrary, takes into account the dependency of previous utterances and their associated emotion labels for the prediction of the emotion label of the current utterance. DialogueGCN models interactions of speakers and it performs slightly better than KET. But it is significantly worse than TodKat. It seems that topics might be more useful in capturing the dialogue context.

Ablation Study The lower half of Table 2 presents the F1 scores with the removal of various components from TodKat. It can be observed that with the removal of the topic component, the performance of TodKat drops consistently across all datasets except IEMOCAP in which we ob-

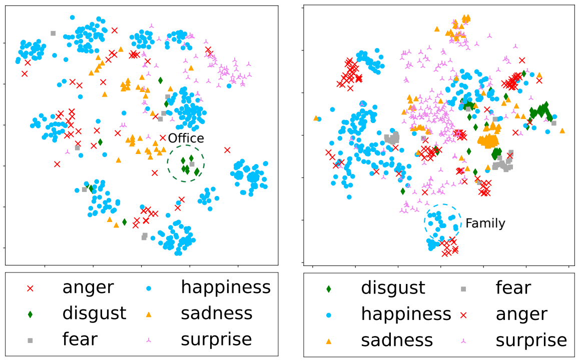
[6] https://huggingface.co/transformers/v2.0.0/examples.html

serve a slight increase in both weighted average F1 and Micro-F1. This might be attributed to the size of the data since IEMOCAP is the smallest dataset evaluated here, and small datasets hinder the model's capability to discover topics. Without using the commonsense knowledge ('$-$KB'), we observe more drastic performance drop compared to all other components, with nearly 10% drop in F1 on EmoryNLP, showing the importance of employing commonsense knowledge for dialogue emotion detection. Comparing two different ways of extracting knowledge from ATOMIC, direct retrieval using SBERT or generation using COMET, we observe mixed results. Overall, the Transformer Encoder-Decoder with a pointer network is a conciliator between the two methods, yielding a balanced performance across the datasets.

Relationships between Topics and Emotions To investigate the effectiveness of the learned topic vectors, we perform t-SNE (Van der Maaten and Hinton, 2008) on the test set to study the relationship between the learned topic vectors and the ground-truth emotion labels. The results on DailyDialog and MELD are illustrated in Figure 3(a) and (b). Latent topic vectors of utterance are used to plot the data points, whose colors indicate their ground-truth emotion labels. We can see that the majority of the topic vectors cluster into polarized groups. Few clusters are bearing a mixture of polarity, possibly due to the background topics such as greetings in the datasets.

Topics can be interpreted using the attention scores of Eq. 4. The top-10 most-attended words are selected as the representative words for each utterance. As in (Dathathri et al., 2020), we construct bag-of-words[7] that represent 141 distinct topics. Given the attended words of an utterance cluster grouped based on their latent topic representations, we label the word collection with the dominant theme name. We refer to the theme names as topics in Figure 3c. It can be observed that utterances associated with Office tend to carry *'disgust'* emotions, while those related to *Family* are prone to be *'happy'*.

We further compute the Spearman's rank-order correlation coefficient to quantitatively verify the relationship between the topic and emotion vectors. For an utterance pair, a similarity score is obtained separately for their corresponding topic vectors as well as their emotion vectors. We then sort the list of emotion vector pairs according to their similarity scores to check to what extent their ranking matches that of topic vector pairs, based on the Spearman's rank-order correlation coefficient. The results are $0.60, 0.58, 0.42$ and 0.54 with p-values $\ll 0.01$ respectively for DailyDialog, MELD, IEMOCAP and EmoryNLP, showing that there is a strong correlation between the clustering of topics and that of emotion labels. IEMOCAP has the lowest correlation score, which is inline with the results in Table 2 that the discovered latent topics did not improve the emotion classification results.



(a) DailyDialog (b) MELD

Topic	Utterances	Emotion
Office	A: How are you doing, Christopher? B: To be honest, I'm really fed up with work at the moment. I need a break! A: Are you doing anything this weekend? B: I have to work on Saturday all day! I really hate my job!	disgust
Family	A: Yeah, I-I heard. I think it's great! Ohh, I'm so happy for you! B: I can't believe you're getting married! C: Yeah. D: Monica and Rachel made out.	happy

(c) Representative utterances and their topics

Figure 3: T-SNE visualization of the learned topic vectors of utterances from the test sets of DailyDialog (subfigure (a)) and MELD (subfigure (b)). Colors indicate the ground-truth emotion label. Neutral utterances are omitted here for clarity. Representative utterances (highlighted in colors) for the topic '*Office*' in DailyDialog and the topic '*Family*' in MELD are shown in subfigure (c).

Impact of Relation Type We investigate the impact of commonsense relation types on the performance of TODKAT. We expand the relation set to five relation types and all nine relation types, respectively. According to (Sap

[7] Word lists and their corresponding theme names are crawled from https://www.enchantedlearning.com/wordlist/.

Dataset	Relation Type	
	$\{sI, sR, oR, sE, oE\}$	All
DailyDialog	0.5718$\downarrow$	0.5664$\downarrow$
MELD	0.6429$\downarrow$	0.6322$\downarrow$
IEMOCAP	0.6163$\uparrow$	0.6073$\downarrow$
EmoryNLP	0.4029$\downarrow$	0.3885$\downarrow$

Table 3: Micro-F1 scores of TODKAT with more commonsense relation types retrieved from ATOMIC included for training. Here, "sE" and "oE" represent *effect of subject* and *effect of object*, respectively. "All" denotes the incorporation of all nine commonsense relation types from ATOMIC.

et al., 2019), there are other relation types including $\{sNeed, sWant, oWant, sEffect, oEffect\}$, which identifies the prerequisites and post conditions of the given event, and $\{sAttr\}$, the "If-Event-Then-Persona" category of relation type that describes how the subject is perceived by others. We calculate the Micro-F1 scores of TODKAT with these two categories of relation types added step by step. From Table 3 we can conclude that the inclusion of two extra relation types or all relation types degrades the F1 scores on almost all datasets. An exception occurs on IEMOCAP where the F1 score rises by 0.5% when adding "sE" and "oE" relations, possibly due to the fact that the dataset is abundant in events. Hence the extra event descriptions offer complementary knowledge to some extent. While on other datasets neither the incorporation of "If-Event-Then-Event" nor the incorporation of "If-Event-Then-Persona" relation types could bring any benefit.

Impact of Attention Mechanism With the knowledge retrieved from ATOMIC or generated from COMET, we are able to infer the possible intentions and reactions of the interlocutors. However, not all knowledge phrases contribute the same to the emotion of the focused utterance. We study the attention mechanism in terms of selecting the relevant knowledge. We show in Table 4 a heat map of the attention scores in Eq. 9 to illustrate how the topic-driven attention could identify the most salient phrase. The utterance '*Oh my God, you're a freak.*' will be erroneously categorized as '*mad*' without using the topic-driven attention (shown in the last row of Table 4). In contrast, the attention mechanism guides the model to attend to the more relevant events and thus predict the correct emotion label.

Dialogue Context	A: Alright, go on.	Neutral
	B: Ok, I have to sleep on the west side because I grew up in California and otherwise the ocean would be on the wrong side.	Neutral
	A: Oh my God, you're a freak.	Joyful
	B: Yeah. How about that.	Neutral
Topic-Driven Attention	A wants to be liked	
	A wants to be accepted	
	A wants to be a freak	
	A will feel satisfied	
	A will feel ashamed	Joyful ✓
	A will feel happy	
	B will feel impressed	
	B will feel disgusted	
	B will feel surprised	
	A: Oh my God, you're a freak.	Mad ✗

Table 4: Illustration of the attention mechanism in Eq. 9 that helps distinguish the retrieved knowledge.

6 Conclusion

We have proposed a Topic-Driven and Knowledge-Aware Transformer model that incorporates topic representation and the commonsense knowledge from ATOMIC for emotion detection in dialogues. A topic-augmented language model based on fine-tuning has been developed for topic extraction. Pointer network and additive attention have been explored for knowledge selection. All the novel components have been integrated into the Transformer Encoder-Decoder structure that enables Seq2Seq prediction. Empirical results demonstrate the effectiveness of the model in topic representation learning and knowledge integration, which have both boosted the performance of emotion detection.

Acknowledgements

The authors would like to thank the anonymous reviewers for insightful comments and helpful suggestions. This work was funded by the EPSRC (grant no. EP/T017112/1, EP/V048597/1). LZ is funded by the Chancellor's International Scholarship at the University of Warwick. YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1). DZ is funded by the National Key Research and Development Program of China (2017YFB1002801) and the National Natural Science Foundation of China (61772132).

References

Meysam Asgari-Chenaghlu, Mohammad-Reza Feizi-Derakhshi, Mohammad-Ali Balafar, and Cina Motamed. 2020. Topicbert: A transformer transfer learning based memory-graph approach for multimodal streaming social media topic detection. *arXiv preprint arXiv:2008.06877*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

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.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. 2019. A discrete cvae for response generation on short-text conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1898–1908.

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

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

Chao-Chun Hsu and Lun-Wei Ku. 2018. SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 27–31, Melbourne, Australia. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8002–8009.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020a. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2020b. Bieru: bidirectional emotional recurrent unit for conversational sentiment analysis. *arXiv preprint arXiv:2006.00492*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International Conference on Machine Learning (ICML)*, pages 1727–1736.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801, New Orleans, Louisiana. Association for Computational Linguistics.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 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*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983, Hong Kong, China. Association for Computational Linguistics.

Mehdi Rezaee and Francis Ferraro. 2020. A discrete variational recurrent topic model without the reparametrization trick. In *Advances in Neural Information Processing Systems*, volume 33, pages 13831–13843.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 3295–3301.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 5456–5463.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, volume 28, pages 3483–3491.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4444–4451.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, volume 27, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in neural information processing systems*, volume 28, pages 2692–2700.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Min Zeng, Yisen Wang, and Yuan Luo. 2019. Dirichlet latent variable hierarchical recurrent encoder-decoder in dialogue generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1267–1272, Hong Kong, China. Association for Computational Linguistics.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5415–5421. International Joint Conferences on Artificial Intelligence Organization.

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

A Appendices

A.1 Settings

We modified the script[8] of language model fine-tuning in the Hugging Face library (Wolf et al., 2020) for the implementation of topic-driven fine-tuning. On each training set, we train the topic model for 3 epochs, with learning rate set to $5e$-5 to prevent overfitting to the low-resource dataset. The classifier is built on the Transformers[9] package in Hugging Face. The language model we employ is RoBERTa (Liu et al., 2019). Each utterance is padded by the `<pad>` token of RoBERTa if it is less than the maximum length of 128. The maximum number of utterances in a dialogue is set to 36, 25, 72 and 25 respectively for DailyDialog (Li et al., 2017) [10], MELD (Poria et al., 2019) [11], IEMOCAP (Busso et al., 2008) [12] and EmoryNLP (Zahiri and Choi, 2018) [13]. Dialogues with shorter lengths are padded with `NULL`. It is worth noting that this step is performed after RoBERTa due to the random noises introduced by RoBERTa. The number of retrieved or generated events from ATOMIC under the relation types '*intentions*' and '*reactions*' is both set to 5, i.e., $K = 5$.

[8] https://huggingface.co/transformers/v2.0.0/examples.html

[9] https://huggingface.co/transformers/

[10] http://yanran.li/dailydialog.html

[11] https://github.com/declare-lab/MELD

[12] https://sail.usc.edu/iemocap/iemocap_release.htm

[13] https://github.com/emorynlp/emotion-detection