Intent-aware Transformer Network for Emotion Recognition in Conversation

Minh-Cong Vo, Guee-Sang Lee, Soo-Hyung Kim, Hyung-Jeong Yang
Department of Artificial Intelligence Convergence, Chonnam National University, Republic of Korea.
congvm.it@gmail.com, gslee@jnu.ac.kr, shkim@jnu.ac.kr, hyungjeong@gmail.com

요 약

Emotion Recognition in Conversation (ERC) is becoming a research trend due to the increase of publicly available conversational data in platforms such as Reddit, Facebook, Twitter, etc. Besides, ERC is also crucial for emotion-aware conversation generation tasks that require understanding the user's emotions. The majority of previous methods focus on exploiting the semantic context in each utterance. However, emotions are not only hidden in semantic context but also stayed in other conversational factors such as topics, speakers' intents, viewpoints, personalities, etc. In this paper, we propose an Intent-aware Transformer Network to exploit the intents of speakers during the conversation for emotion recognition. We firstly build a ToD-Roberta model to extract the intent information of each utterance. Then, a Transformer Network inputs a sequence of extracted intent features to infer the emotion for each utterance. The model has experimented on the MELD dataset. The results show that intent information can be helpful to distinguish emotion categories.

Keywords: Emotion Recognition, Intent-aware, Deep Learning, Transformer, Intent Recognition

1. Introduction

In the past few years, Emotion Recognition in Conversation (ERC) gained much attention from the NLP community [1, 2]. A large amount of online conversations provides a vast opportunity to develop an AI system for emotion recognition, which is essential for specific applications that require user's emotion understanding, such as conversational agents or chatbot [3, 4]. Unlike vanilla emotion recognition of utterances, ERC requires contextual semantic modelling of the individual utterances in a conversation. This context can be attributed to the previous utterances and relies on the temporal sequence of utterances. However, it is challenging to capture the contextual semantics described in an utterance. For example, the emotion in the sentence "Yeah!" can be either happy or sad, relying on the speaker's intent or understanding. Previous methods address this problem by utilizing the dialogue context to improve the utterance representation [5, 6, 7, 8], where the recurrent and attention neural network implemented to capture the influences among utterances in a conversation.

Despite the considerable effort by the methods mentioned above, ERC is still a challenging task because the meaning of each utterance varies based on the expectation or experience of the speaker during the conversation. Moreover, emotions are usually hidden in various conversational factors such as topics, speakers' intents, viewpoints, personalities, etc. In this paper, we propose an Intent-based Transformer Network to exploit the intents of speakers during the conversation for emotion recognition. We firstly build a ToD-Roberta model to extract the intent information of each utterance. Then, a Transformer Network infers the emotion from the sequence of extracted intent information.

2. Related Works

2.1. Dialogue Emotion Detection

The majority of prior research exploited the importance of dialogue context in dialogue emotion detection. Majumder et al. [9] by using Gated Recurrent Unit (GRU) to capture the global context. Jiao et al. [5] proposed a hierarchical neural network model that comprises two GRUs for the modelling of the meaning in each utterances and among utterances respectively. Zhang et al. [6] utilized a Graph Convolutional Network (GCN) to explicitly model the emotional dependencies on context and speakers, Ghosal et al. [10] extended the prior work [9] by taking into account the intra-speaker dependency and relative position of the target and context within a conversation. Some authors have also suggested exploring memory networks [11] to allow bidirectional influence between utterances. [12] enriched utterances with concept representations captured from the ConceptNet [13]. While the existing works have concentrated on textual context, Ghosal et al. [8] proposed COSMIC, which exploited ATOMIC

[14] for the acquisition of commonsense knowledge.

2.2. Intent Classification

Intent recognition is the task of taking an input utterance and classifying it based on what the speaker wants to achieve. Intent recognition becomes an essential component of chatbots and finds use in customer support, sales conversions, and many other areas. The intent classifier is expected to map an input utterance to the correct intent and detect when the utterance is unrelated to any of the pre-defined intents, referred to as out-of-scope (OOS) samples. Previous studies considered this task as a standard text classification task [15, 16]. To perform OOS sample detection, these methods add an extra out-of-class class or use threshold rejection techniques on the probability outputs for each class [15], or reconstruction errors [17]. Other approaches used transfer learning techniques. Wu et al. [18] proposed a customized version on BERT trained on a large NLP dataset and then fine-tune their model for intent classification task.

3. Proposed Method

In this section, we present our Intent-aware Transformer Network shown in Fig. 1.

3.1. Problem Setup

Given the transcript of a conversation, the ERC task aims to detect the emotion of each utterance from several pre-defined emotions. Formally, given the input sequence of N number of utterances $u_1, u_2, ..., u_N$ which is annotated with a sequence of emotion labels $y_1, y_2, ..., y_N$ where each utterance $u_i = [w_{i,1}, w_{i,2}, ..., w_{i,T}]$ consists of T words $w_{i,j}$ and spoken by party p_i. The task consecutively inputs an utterance u_i and predict the emotion label y_i based on the preceding utterances and their associated predicted emotion labels.

3.2. Intent Feature Extraction

In this section, we leverage ToD-Bert [18] to perform intent feature extraction. We change the language model to Roberta [19] for better performance instead of BERT [20] as the original model. We then call the modified version as ToD-Roberta. Afterwards, the Tod-Roberta is fine-tuned on The out-of-scope intent dataset (OOS) [15]. The OOS intent dataset is one of the largest annotated intent datasets, consisting of more than 20,000 samples. This dataset is split into three parts for the train, validation, and test sets. The labels cover 151 intent classes over ten domains for 150 in-scope intent and one out-of-scope intent. We remove the classification layers of the modified model for latent vector retrieval.

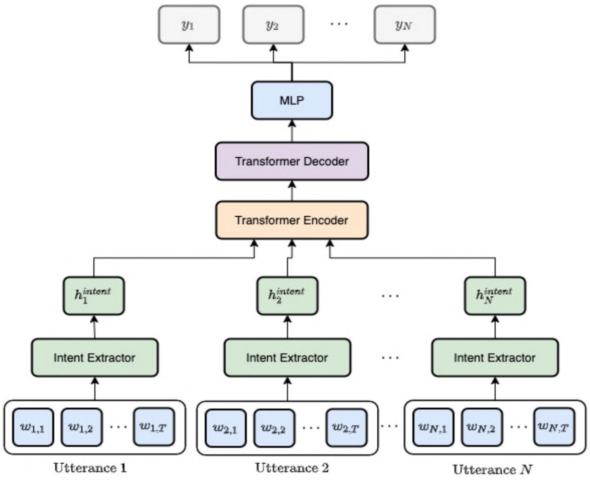

Fig 1. Our proposed method contains two modules. We first extract the intent feature h_i^{intent} by passing each utterance through the intent extractor. Then we input a sequence of intent features to Transformer Network and MLP to infer emotion y_i.

3.3. Transformer-based Classifier

In this section, we use a Transformer [21] Network to map an intent features sequence to an emotion label sequence, which allows capturing the transitional patterns between emotions and the contextual semantic. Each intent feature h_i^{intent} is converted to the [CLS] representation and infer the emotion label. In this experiment, we focus on solving the ERC task in an auto-regressive way by using a masking scheme in the self-attention layer of the encoder to predict emotions, which makes sure that only the past utterances are presented to the encoder. This strategy prevents the leaking of information from future utterances, which suits a real-world scenario. For the decoder, the output of the previous decoder block inputs to the self-attention layer as a query. The training loss is the negative log-likelihood expressed as below:

$$L = -\sum_{i=1}^{26} \log p_\theta (y_n | h_{\le n}^{intent}, y_{<n})$$

Where θ denotes the trainable parameters.

4. Experimental Results

4.1. MELD Dataset

Multimodal EmotionLines Dataset (MELD) [22] has been created by enhancing and extending EmotionLines dataset [1]. MELD consists of approximately 1400 dialogues with more than 13000 utterances from the Friends TV series. Multiple speakers participated in the dialogues. Each utterance in dialogue has been annotated by one of these seven emotions, such as

Disgust, Anger, Sadness, Joy, Surprise, Neutral and Fear. MELD also has sentiment (positive, negative and neutral) annotation for each utterance. In this paper, we only focus on the seven-emotion classification.

4.2. Implementation

Our method is implemented by using Pytorch [24] framework. For the intent information extraction, we first train the TOD-RoBERTA model on OSS Intent Dataset. Then we remove the last layer for classification for intent latent vector retrieval. For emotion recognition, Our Transformer Network is trained and optimized using Adam optimizer [25] with learning set to 5e-5 to prevent overfitting. In addition, each utterance is padded by the [PAD] token of Roberta if its length is less than 128. The model experiments on a desktop PC with AMD Ryzen 7 2700X equipped with an NVIDIA GTX 2080Ti GPU processor.

Table 1. Results comparison on the MELD dataset

Models	Weighted Avg-F1	Micro-F1
HiGRU [5]	0.5681	0.5452
Dialogue [10]	0.5837	0.5617
KET [7]	0.5818	-
COSMIC [23]	0.6521	-
Our (w/ Intent)	**0.6650**	**0.6464**
Our (w/o Intent)	0.6421	0.6215

4.3. Results on MELD Dataset

This section describes our results on the test set of the MELD dataset. Table 1 summarizes the performance of the proposed model compared with previous models, in which our model outperforms previous methods on both weighted and micro average F1 metrics. To motivate the importance of intent information, we run our model on the MELD dataset with and without the intent information. The results of ablations experiments are also summarized in Table 1.

5. Conclusion

This paper proposes an Intent-based Transformer Network to exploit the intents of speakers during the conversation for emotion recognition. We first build a ToD-Roberta model trained on the Intent Recognition dataset to extract the intent information for each utterance. Then, a Transformer Network uses that extracted intent information to infer the emotion. Finally, the model has experimented on the MELD dataset. The results show that intent information can be helpful to distinguish emotion. Furthermore, the exploitation of the dependencies between hidden conversational variables and emotions can improve the performance of this ERC task.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2020R1A4A1019191) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058). The corresponding author is Guee-Sang Lee.

References

[1] Sheng-Yeh Chen et al. “Emotionlines: An emotion corpus of multi-party conversations”. In: arXiv preprint arXiv:1802.08379 (2018).

[2] Hao Zhou et al. “Emotional chatting machine: Emotional conversation generation with internal and external memory”. In: Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[3] Chao-Chun Hsu and Lun-Wei Ku. “Socialnlp 2018 emotionx challenge overview: Recognizing emotions in dialogues”. In: Proceedings of the sixth international workshop on natural language processing for social media. 2018, pp. 27 – 31.

[4] Jie Cao et al. “Observing dialogue in therapy: Categorizing and forecasting behavioral codes”. In: arXivpreprint arXiv:1907.00326 (2019).

[5] Wenxiang Jiao et al. “Higru: Hierarchical gated recurrent units for utterance-level emotion recognition”. In: arXiv preprint arXiv:1904.04446 (2019).

[6] Dong Zhang et al. “Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations.” In: IJCAI. 2019, pp. 5415 – 5421.

[7] Peixiang Zhong, Di Wang, and Chunyan Miao. “Knowledge-enriched transformer for emotion detection in textual conversations”. In: arXiv preprint arXiv:1909.10681 (2019).

[8] Deepanway Ghosal et al. “COSMIC: COmmonSense knowledge for eMotion Identification in Conversations”. In: arXiv preprint arXiv:2010.02795 (2020).

[9] Navonil Majumder et al. “Dialoguernn: An attentive rnn for emotion detection in conversations”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 01. 2019, pp. 6818-6825.

[10] Deepanway Ghosal et al. “Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation”. In: arXiv preprint arXiv:1908.11540 (2019).

[11] Wenxiang Jiao, Michael Lyu, and Irwin King.

"Realtime emotion recognition via attention gated hierarchical memory network". In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. 05. 2020, pp. 8002 - 8009.

[12] Peixiang Zhong, Di Wang, and Chunyan Miao. "Knowledge-enriched transformer for emotion detection in textual conversations". In: arXiv preprint arXiv:1909.10681 (2019).

[13] Robyn Speer, Joshua Chin, and Catherine Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge". In: Thirty-first AAAI conference on artificial intelligence. 2017.

[14] Maarten Sap et al. "Atomic: An atlas of machine commonsense for if-then reasoning". In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 01. 2019, pp. 3027 - 3035.

[15] Stefan Larson et al. "An evaluation dataset for intent classification and out-of-scope prediction". In: arXiv preprint arXiv:1909.02027 (2019).

[16] Inigo Casanueva et al. "Efficient intent detection with dual sentence encoders". In: arXiv preprint arXiv:2003.04807 (2020).

[17] Seonghan Ryu et al. "Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems". In: Pattern Recognition Letters 88 (2017), pp. 26 - 32.

[18] Chien-Sheng Wu et al. "TOD-BERT: pre-trained natural language understanding for task-oriented dialogue". In: arXiv preprint arXiv:2004.06871 (2020).

[19] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: arXiv preprint arXiv:1907.11692 (2019).

[20] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: arXiv preprint arXiv:1810.04805 (2018).

[21] Ashish Vaswani et al. "Attention is all you need".In: Advances in neural information processing systems. 2017, pp. 5998 - 6008.

[22] Soujanya Poria et al. "Meld: A multimodal multi-party dataset for emotion recognition in conversations". In: arXiv preprint arXiv:1810.02508 (2018).

[23] Deepanway Ghosal et al. "COSMIC: COmmonSense knowledge for eMotion Identification in Conversations". In: arXiv preprint arXiv:2010.02795 (2020).

[24] Adam Paszke et al. "Automatic differentiation in pytorch". In: (2017).

[25] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: arXiv preprint arXiv:1412.6980 (2014).