# CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation

**Anonymous ACL submission**

## Abstract

As the use of interactive machines grow, the task of Emotion Recognition in Conversation (ERC) became more important. If the machine generated sentences reflect emotion, more human-like sympathetic conversations are possible. Since emotion recognition in conversation is inaccurate if the previous utterances are not taken into account, many studies reflect the dialogue context to improve the performances. We introduce CoMPM, a context embedding module (CoM) combined with a pre-trained memory module (PM) that tracks memory of the speaker's previous utterances within the context, and show that the pre-trained memory significantly improves the final accuracy of emotion recognition. We achieve competitive performance with previous methods on English datasets (MELD, EmoryNLP, IEMOCAP, DailyDailog), and achieve good performance with small data sets. In addition, our method shows that it can be extended to other languages because structured knowledge is not required unlike existing methods.

## 1 Introduction

As the number of applications such as interactive chatbots or social media that are used by many users has recently increased dramatically, Emotion Recognition in Conversation (ERC) plays a more important role in natural language processing, and as a proof, a lot of research (Poria et al., 2019; Zhang et al., 2019; Ghosal et al., 2020; Jiao et al., 2020) has been conducted on the task.

The ERC module increases the quality of empathetic conversations with the users and can be utilized when sending tailored push messages to the users (Shin et al., 2019; Zandie and Mahoor, 2020; Lin et al., 2020). In addition, emotion recognition can be effectively used for opinion mining, recommender systems and healthcare systems where it can improve the service qualities by providing personalized results. As these interactive machines increase, the ERC module plays an increasingly important role.

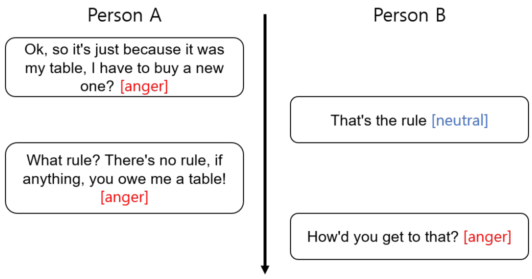


Figure 1: An example of MELD dataset

Figure 1 is an example of a conversation in which two speakers are angry at each other. The emotion of speaker B's utterance (*"How'd you get to that?"*) is *angry*. If the system does not take into account previous utterances, it is difficult to properly recognize emotions. Like the previous studies (Ghosal et al., 2020), we show that the utterance-level emotion recognition, which do not consider the previous utterance, have limitations and experiments result in poor performances.

Therefore, recent studies are attempting to recognize emotions while taking into account the previous utterances. Representatively, DialogueRNN (Majumder et al., 2019) recognizes the present emotion by tracking context from the previous utterances and the speaker's emotion. AGHMN (Jiao et al., 2020) considers the previous utterances through memory summarizing using GRU with attention.

COSMIC (Ghosal et al., 2020) and KET (Zhong et al., 2019) use external knowledge to improve the ERC performance. COSMIC improves the performance of emotion recognition by extracting commonsense knowledge of the previous utterances. Commonsense knowledge feature is extracted and leveraged with COMET (Bosselut et al., 2019) trained with ATOMIC (The Atlas of Machine Commonsense) (Sap et al., 2019). ATOMIC has 9 sentence relation types with inferential if-then commonsense knowledge expressed in text. KET is used as external knowledge based on ConceptNet (Speer et al., 2017) and emotion lexicon NRC_VAD (Mohammad, 2018) as the commonsense knowledge. ConceptNet is a knowledge graph that connects words and phrases in natural language using labeled edges. NRC_VAD Lexicon has human ratings of valence, arousal, and dominance for more than 20,000 English words. However, this external knowledge is often only available in English. In order to utilize the previous methods in languages of other countries, it is expensive and difficult to utilize because external knowledge data must be newly constructed. In recent NLP studies, due to the effectiveness of the pre-trained language model, it has already been developed in many countries. Additionally, Petroni et al. (2019) introduces that the language models can be used as knowledge bases and have many advantages over the structured knowledge bases. Based on these studies, we introduce an approach using pre-trained memory tracking of previous utterances that can be used regardless of the speaker's language.

CoMPM, introduced in this paper, is composed of two modules that take into account previous utterances in dialogue. (1) The first is a context embedding module (CoM) that reflects all previous utterances as context. CoM is an auto-regressive model that predicts the current emotion through attention between the previous utterances of the conversation and the current utterance. (2) The second is a pre-trained memory module (PM) that extracts memory from utterances. We use the output of the pre-trained language model as the memory embedding where the utterances are passed into the language model. We use the PM to help predict the emotion of the speaker by taking into account the speaker's linguistic preferences and characteristics.

We experiment on 4 different English ERC datasets. Multi-party datasets are MELD (Poria et al., 2019) and EmoryNLP (Zahiri and Choi, 2018), and dyadic datasets are IEMOCAP (Busso et al., 2008) and DailyDialog (Li et al., 2017). CoMPM achieves the first or second performance according to the evaluation metric compared to all previous systems. We performed an ablation study on each module to show that the proposed approach is effective. Further experiments also show that our approach can be used in other languages or small data similar to the limited service environment.

## 2 Related Work

Ekman (Ekman, 1992) constructs taxonomy of six common emotion (Joy, Sadness, Fear, Anger, Surprise, and Disgust) from human facial expressions. In addition, Ekman explains that multi-modal view is important for multiple emotions recognition. The multi-modal data such as MELD and IEMOCAP are some of the available standard datasets for emotion recognition and they are composed of text, speech and vision-based data. Datcu and Rothkrantz (2014) uses speech and visual information to recognize emotions, and (Alm et al., 2005) attempts to recognize emotions based on text information. MELD and ICON (Hazarika et al., 2018a) show that the more multi-modal information is used, the better the performance and the text information plays the most important role. Multi-modal information is not always given in most social media, especially in chatbot systems where they are mainly composed of text-based systems. In this work, we design and introduce a text-based emotion recognition system using neural networks.

In the previous studies, such as Hazarika et al. (2018b); Zadeh et al. (2017); Majumder et al. (2019), most works focused on dyadic-party conversation. However, as the multi-party conversation datasets including MELD and EmoryNLP have become available, a lot of recent research are being conducted on multi-party dialogues such as Zhang et al. (2019); Jiao et al. (2020); Ghosal et al. (2020). In general, the multi-party conversations have higher speaker dependency than the dyadic-party dialogues, therefore have more conditions to consider and result in poor performance.

Zhou et al. (2018); Zhang et al. (2018a) shows that commonsense knowledge is important for understanding conversations and generating appropriate responses. Liu et al. (2020) reports that the lack of external knowledge makes it difficult to classify implicit emotions from the conversation history. EDA (Bothe et al., 2020) expands the multi-modal emotion datasets by extracting dialog acts from MELD and IEMOCAP and finds out that there is a correlation between dialogue acts and emotion labels.

## 3 Approach

### 3.1 Problem Statement

In a conversation, $M$ sequential utterances are given as $[(u_1, p_{u_1}), (u_2, p_{u_2}), ..., (u_M, p_{u_M})]$. $u_i$ is the utterance which the speaker $p_{u_i}$ uttered, where $p_{u_i}$ is one of the conversation participants. While $p_{u_i}$ and $p_{u_j}$ $(i \neq j)$ can be the same speaker, the minimum number of the unique conversation participants should be 2 or more. The ERC is a task of predicting the emotion $e_t$ of $u_t$, the utterance of the $t$-th turn, given the previous utterances $h_t = \{u_1, ..., u_{t-1}\}$. Emotions are labeled as one of the predefined classes depending on the dataset, and the emotions we experimented with are either 6 or 7. We also experimented with a sentiment classification dataset which provides sentiment labels consisting of positive, negative and neutral.

### 3.2 Model Overview

Figure 2 shows an overview of our model. Our ERC neural network model is composed of two modules. The first is CoM which catches the underlying effect of all previous utterances on the current speaker's emotions. Therefore, we propose a context model to handle the relationship between the current and the previous utterances. The second one is PM that leverages only the speaker's previous utterances, through which we want to reflect the speaker specific preferences and characteristics.

### 3.3 CoM: Context Embedding Module

The context embedding module predicts $e_t$ by considering all of the utterances before the $t$-th turn as the dialogue context. The example in Figure 2 shows how the model predicts the emotion of $u_6$ uttered by $s_A$, given a conversation of three participants ($s_A$, $s_B$, $s_C$). The previous utterances are $h_6 = \{u_1, \cdots u_5\}$ and $e_6$ is predicted while considering the relationship between $u_6$ and $h_6$.

We consider multi-party conversations where 2 or more speakers are involved. A special token $<s_{\mathbb{P}}>$ is introduced to distinguish participants in the conversation and to handle the speaker's dependency where $\mathbb{P}$ is the set of participants. In other words, the same special token appears before the utterances of the same speaker.

The context model operates auto-regressively and follows the causal decoder architecture where only the left context is used to predict the next word. Therefore, when the model predicts $e_t$, there is no effect of the future utterances. In many natural language processing tasks, the effectiveness of the pre-trained language model has been proven, and we also set the initial state of the model to GPT2 (Radford et al., 2018). GPT2 is an unsupervised pre-trained model with large-scale open-domain corpora of unlabeled text.

We use the embedding of the special token <cls> to predict emotion. The <cls> token is concatenated at the end of the input and the output of the context model is as follows:

$$\boldsymbol{c}_t = \text{Context-Model}(\mathbb{P}_{:t-1}, h_t, u_t, <\text{cls}>) \quad (1)$$

where $\mathbb{P}_{:t-1}$ is the set of speakers in the previous turns. $\boldsymbol{c}_t \in \mathbb{R}^{1 \times h_c}$ and $h_c$ is the dimension of Context-Model.

### 3.4 PM: Pre-trained Memory Module

External knowledge is known to play an important role in understanding conversation. Pre-trained language models can be trained on numerous corpora and be used as an external knowledge base. We utilize the pre-trained embedding of the speaker's previous utterances to compute and predict the emotion of the current utterance $u_t$. If the speaker has never appeared before the current turn, the result of the pre-trained memory is considered a zero vector.

To extract utterance-level embeddings, a pre-trained language model with a bidirectional encoder structure is used. We use the distilled version of the RoBERTa (Liu et al., 2019) model, distilRoBERTa. DistilRoBERTa is trained with the same training procedure as distilBERT (Sanh et al., 2019), and the number of parameters is 65.6% of RoBERTa. We used distilRoBERTa-base since no significant difference in performance was found using other pre-trained language models.

Since <cls> is mostly used for the task of classifying sentences, we use the embedding output of the <cls> token as a vector representing the utterance as follows:

$$\boldsymbol{k}_i = \text{Memory-Encoder}(<\text{cls}>, u_i) \quad (2)$$

where $p_{u_i} = p_S$, S is the speaker of the current utterance. $\boldsymbol{k}_i \in \mathbb{R}^{1 \times h_k}$ and $h_k$ is the dimension of Memory-Encoder.

### 3.5 CoMPM: Combination of CoM and PM

We combine CoM and PM to predict the speaker's emotion. In many dialogue systems (Zhang et al., 2018b; Ma et al., 2019), it is known that utterances close to the current turn are important for response. Therefore, we assume that utterances close to the current utterance will be important in emotional

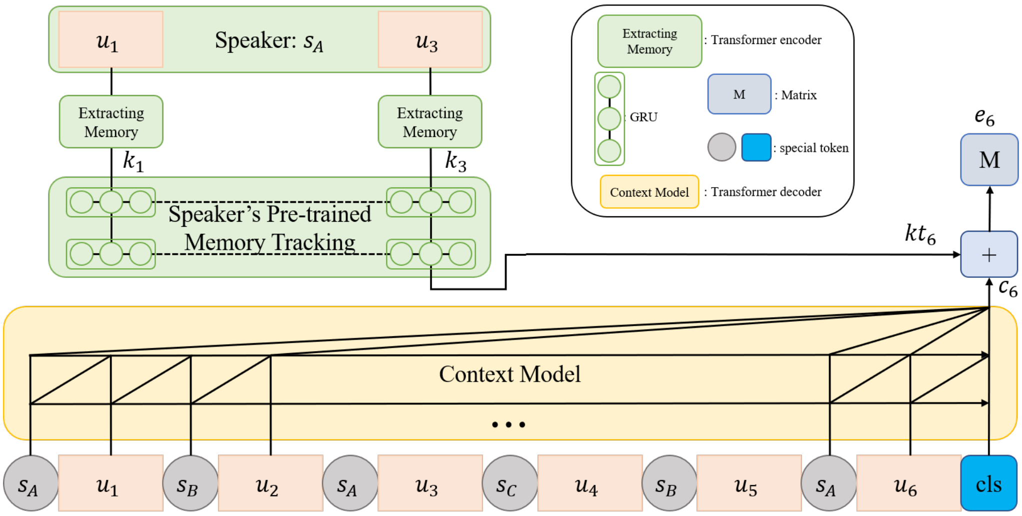


Figure 2: Our model consists of two modules: a context embedding module and a pre-trained memory module. The figure shows an example of predicting emotion of $u_6$, from a 6-turn dialogue context. A, B, and C refer to the participants in the conversation, where $s_A = p_{u_1} = p_{u_3} = p_{u_6}, s_B = p_{u_2} = p_{u_5}, s_C = p_{u_3}$. $\mathbf{M}$ is a linear matrix

recognition. To confirm this assumption, two methods are presented in this section for combining $\boldsymbol{c}_t$ and $\boldsymbol{k}_i$ as a result of Equation 1, 2.

#### 3.5.1 Tracking Method

The first is a $\boldsymbol{k}_i$ tracking method using GRU. The tracking method assumes that the importance of all previous speaker utterances to the current emotion is not equal and varies with the distance of the current utterance. In other words, since the flow of conversation changes as it progresses, the effect on emotion may differ depending on the distance from the current utterance. We track and capture the sequential position information of $\boldsymbol{k}_i$ using a unidirectional GRU:

$$\boldsymbol{kt}_t = \mathrm{GRU}(\boldsymbol{k}_{i_1}, \boldsymbol{k}_{i_2}, ..., \boldsymbol{k}_{i_n}) \tag{3}$$

where $t$ is the turn index of the current utterance, $n$ is the number of previous utterances of the speaker, and $i_s$ ($s = 1, 2, ..., n$) is each turn uttered. $\boldsymbol{kt}_t \in \mathbb{R}^{1 \times h_c}$ is the output of $\boldsymbol{k}_{i_n}$ and as a result, the knowledge of distant utterance is diluted and the effect on the current utterance is reduced.

GRU is composed of 2-layers, the dimension of the output vector is $h_c$, and the dropout is set to 0.3 during training. Finally, the output vector $\boldsymbol{o}_t$ is obtained by adding $\boldsymbol{kt}_t$ and $\boldsymbol{c}_t$ in Equation 4.

$$\boldsymbol{o}_t = \boldsymbol{c}_t + \boldsymbol{kt}_t \tag{4}$$

#### 3.5.2 Attention Method

The attention method determines the importance of the previous utterances with an attention score instead of a distance based dilution. The attention value is obtained through the similarity between the context-reflected vector $\boldsymbol{c}_t$ and the pre-trained memory vector $\boldsymbol{k}_{i_s}$. Considering that the two vectors are not of the same dimension, we calculate the attention score using a projection matrix $\mathbf{W}$ as follows:

$$a_{i_s} = \boldsymbol{k}_{i_s} \mathbf{W} \boldsymbol{c}_t \tag{5}$$

$$w_{i_{1:n}} = \mathrm{softmax}(a_{i_{1:n}}) \tag{6}$$

$$\boldsymbol{o}_t = \boldsymbol{c}_t + \sum_{s=1}^{n} w_{i_s} \boldsymbol{k}_{i_s} \tag{7}$$

where $W \in \mathbb{R}^{h_k \times h_c}$ is a matrix for calculating the association between $\boldsymbol{k}_{i_s}$ and $\boldsymbol{c}_t$, and the weights $w_{i_{1:n}}$ are obtained from Equation 5, 6. Unlike the tracking method, $\boldsymbol{o}_t$ is calculated as Equation 7 by weighted sum of all pre-trained memory and adding $\boldsymbol{c}_t$.

#### 3.5.3 Emotion Prediction

Softmax is applied to the vector multiplied by $\boldsymbol{o}_t$ and the linear matrix $\mathbf{M} \in \mathbb{R}^{h_e \times h_c}$ to obtain the probability distribution of emotion classes, where $h_e$ is the number of emotion classes. $e_t$ is the predicted emotion class that corresponds to the index of the largest probability from the emotion class distribution.

$$e_t = \underset{e}{\mathrm{argmax}}\, \mathrm{softmax}(\mathbf{M}(\boldsymbol{o}_t)) \tag{8}$$

The objective is to minimize the cross entropy loss so that $e_t$ is the same as the ground truth emotional label.

# 4 Experiments

## 4.1 Dataset

We experiment on four benchmark datasets. MELD (Poria et al., 2019) and EmoryNLP (Zahiri and Choi, 2018) are multi-party datasets, while IEMOCAP (Busso et al., 2008) and DailyDialog (Li et al., 2017) are dyadic-party datasets. The statistics of the dataset are shown in Table 1.

IEMOCAP is a dataset involving 10 speakers, and each conversation involves 2 speakers and the emotion-inventory is given as "happy, sad, angry, excited, frustrated and neutral". The train and development dataset is a conversation involving the previous eight speakers, and the train and development are divided into random splits at a ratio of 9:1. The test dataset is a conversation involving two later speakers.

DailyDialog is a dataset of daily conversations between two speakers and the emotion-inventory is given as "anger, disgust, fear, joy, surprise, sadness and neutral". Since more than 82% of the data are tagged as neutral, neutral emotions are excluded when evaluating systems with Micro-F1 as did in the previous studies.

MELD is a dataset based on Friends TV show and provides two taxonomy: emotion and sentiment. MELD's emotion-inventory is given as "anger, disgust, sadness, joy, surprise, fear and neutrality" following Ekman (Ekman, 1992) and sentiment-inventory is given as "positive, negative and neutral".

EmoryNLP, like MELD, is also a dataset based on Friends TV show, but the emotion-inventory is given as "joyful, peaceful, powerful, scared, mad, sad and neutral". Sentiment labels are not provided, but sentiment classes can be grouped as follows: positive: {joyful, peaceful, powerful}, negative: {scared, mad, sad}, neutral: {neutral}

## 4.2 Training Setup

In CoMPM, CoM uses a pre-trained GPT2-medium as the initial state and PM uses a pre-trained distilRoBERTa as the initial state. We use the pre-trained model from the huggingface library [1]. The optimizer is AdamW and the learning rate is 1e-5 as an initial value. The learning rate scheduler used for training is *get_linear_schedule_with_warmup*, and the maximum value of 10 is used for the gradient clipping. We select the model with the best performance on the validation set. All experiments are conducted on one V100 GPU with 32GB memory.

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

## 4.3 Previous Method

We show that the proposed approach is effective by comparing it with various baselines and the state-of-the-art methods.

**CNN** (Kim, 2014) is a convolutional neural network model using pre-trained GloVe embeddings.

**ICON** (Hazarika et al., 2018a) is composed of GRUs as a model that predicts emotions by hierarchically integrating self- and inter-speaker emotional influences into the global memories.

**KET** (Zhong et al., 2019) is a Knowledge Enriched Transformer that reflects contextual utterances with a hierarchical self-attention and leverages external commonsense knowledge by using a context-aware affective graph attention mechanism.

**ConGCN** (Zhang et al., 2019) is a conversational graph-based convolutional neural network that considers each utterance and speaker as nodes. This model recognizes emotions by expressing context- and speaker-sensitive dependency using the nodes and the edges of the graph.

**DialogueRNN** (Majumder et al., 2019) uses a GRU network to keep track of the individual party states in the conversation to predict emotions. This model assumes that there are three factors in emotion prediction: the speaker, the context from the preceding utterances and the emotion of the preceding utterances. Also, Ghosal et al. (2020) shows the performance of **RoBERTa+DialogueRNN** when the vectors of the tokens are extracted with a pre-trained RoBERTa.

**BERT DCR-Net** (Qin et al., 2020) proposes Deep Co-Interactive Relation Network (DCR-Net) and integrates mutual knowledge by modeling the relation and the interaction between two tasks as a co-interactive relation layer in a multi-task.

**AGHMN** (Jiao et al., 2020) (Attention Gated Hierarchical Memory Network) is composed of (1) a hierarchical memory network through BiGRU and (2) an attention GRU (AGRU) using attention weights to predict emotion.

**RGAT+P** (Ishiwatari et al., 2020) (relational graph attention networks) proposes relational position encodings with sequential information reflect-

| Dataset | dialogues | | | utterance | | | classes | Evaluation Metrics |
|---|---|---|---|---|---|---|---|---|
| | train | dev | test | train | dev | test | | |
| IEMOCAP | 108 | 12 | 31 | 5163 | 647 | 1623 | 6 | weighted avg F1 |
| DailyDialog | 11118 | 1000 | 1000 | 87170 | 8069 | 7740 | 7(6) | Macro F1 & Micro F1 |
| MELD | 1038 | 114 | 280 | 9989 | 1109 | 2610 | 3, 7 | weighted avg F1 |
| EmoryNLP | 713 | 99 | 85 | 9934 | 1344 | 1328 | 3, 7 | weighted avg F1 |

Table 1: Statistics and descriptions for the four datasets. DailyDialog uses 7 classes for training, but we measure Macro-F1 for only 6 classes excluding neutral. MELD and EmoryNLP are used to measure weighted avg F1 for both emotion (7) and sentiment (3) classes.

ing the relational graph structure, which shows that both the speaker dependency and the sequential information can be captured.

**COSMIC** (Ghosal et al., 2020) incorporates different elements of commonsense such as mental states, events and causal relations, and learns the relations between participants in the conversation. This model uses pre-trained RoBERTa as a feature extractor and leverages COMET trained with ATOMIC as the commonsense knowledge.

### 4.4 Result and Analysis

Table 2 shows the performance of the previous methods and our models. CoM used alone does not leverage PM and predicts emotions by only considering the dialogue context. PM, if used alone, does not consider the context and predicts emotions only with the utterance of the current turn. CoMPM is a model that combines CoM and PM with the tracking method (3.5.1), and CoMPM-A is a model that combines the two modules with the attention method (3.5.2). CoMPM(s) is a model in which PM is trained from scratch.

Compared to other models using external knowledge, CoMPM achieves effective performance without the need for new training and other data. In other words, we can infer that the pre-trained language model is more effective as external knowledge than ATOMIC (Sap et al., 2019), ConceptNet (Speer et al., 2017) or NRC_VAD (Mohammad, 2018). In addition, CoM, RoBERTa+DialogueRNN, BERT DCR-Net, and RGAT+P use pre-trained models as an initial state or feature extractor, but their performance is worse than CoMPM. Experimental results show that CoMPM is more effective than simply using a pre-trained language model as a backbone or feature extractor.

When comparing the differences in performance between CoMPM and CoM, the effect of PM can be validated, and when compared with PM, the effect of CoM can be confirmed. CoM and PM each show inferior performance compared to the baselines, but we achieved higher performance by integrating the two and confirmed that each module is an important factor. In addition, PM does not consider the context, so the performance is worse than CoM and the performance gap is even greater in IEMOCAP datasets with longer average conversation turns.

The difference in performance between CoMPM and CoMPM-A comes from the difference in the method of combining the pre-trained memory. We find that the tracking method is more effective than the attention method in predicting emotions. Since the tracking method uses unidirectional GRU, the knowledge extracted from distant speaker utterances is diluted. On the other hand, the attention method determines the weight through attention between all of the speaker's utterances and the current utterance. Therefore, information about sequential and position is not reflected. We experimentally find that the sequential and position information can be an important factor, and that the proximal utterances of the speaker have a higher influence on the emotion classification, which is more prominent on the IEMOCAP data with a longer average turns of conversation.

We confirm the effect of PM structure in the model through the performance of CoMPM(s). If PM is randomly initialized and trained, the performance deteriorates because PM does not play the role of a pre-trained memory. CoMPM(s) slightly shows better performance than CoM, but slightly inferior to CoMPM. That is, PM used in CoMPM(s) cannot be considered as a pre-trained memory, but it is used to extract and utilize features from previous utterances of the speaker. Feature vectors extracted with PM are trained to help predict emotions by reflecting the speaker specific personality and characteristics.

| Models | IEMOCAP | DailyDialog | | MELD | | EmoryNLP | |
|---|---|---|---|---|---|---|---|
| | W-Avg F1 | Macro F1 | Micro F1 | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) |
| CNN | 52.04 | 36.87 | 50.32 | 64.25 | 55.02 | 38.05 | 32.59 |
| ICON | 58.54 | - | - | - | - | - | - |
| KET | 59.56 | - | 53.37 | - | 58.18 | - | 34.39 |
| DialogueRNN | 62.57 | 41.8 | 55.95 | 66.1 | 57.03 | 48.93 | 31.7 |
| RoBERTa DialougeRNN | 64.76 | 49.65 | 57.32 | 72.14 | 63.61 | 55.36 | 37.44 |
| BERT DCR-Net | - | 48.9 | - | - | - | - | - |
| ConGCN | - | - | - | - | 59.4 | - | - |
| AGHMN | - | - | - | - | 59.03 | - | - |
| RGAT+P | 65.22 | - | 54.31 | - | 60.91 | - | 34.42 |
| COSMIC | 65.28 | 51.05 | 58.48 | 73.2 | 65.21 | 56.51 | 38.11 |
| CoMPM | **65.79** | **53.14** | **59.63** | **73.6** | **64.62** | **58.35** | **37.44** |
| CoM | 62.44 | 49.76 | 54.17 | 70.95 | 63.65 | 57.67 | 36.34 |
| PM | 50.37 | 46.73 | 50.48 | 70.36 | 61.5 | 54.74 | 35.5 |
| CoMPM(s) | 63.29 | 51.36 | 56.73 | 72.04 | 63.61 | 57.69 | 36.46 |
| CoMPM-A | 62.7 | 51.11 | 56.01 | 71.69 | 63.49 | 57.86 | 35.76 |

Table 2: Comparison of our models with various previous models and the results on 4 datasets. Our models are trained 3 times for each experiment and the average of the scores is evaluated (same in other tables). Test performance is measured by the model with the best score in the validation dataset. CoMPM, in bold text, is our final results.

| Conversation | speaker | utterance | pred | label |
|---|---|---|---|---|
| #1 | A | Eh..., I don't, I don't know. | neutral | sadness |
| | B | What? | surprise | neutral |
| | ⋮ | ⋮ | ⋮ | ⋮ |
| | C | Good one. Actually, ah, Terry wants you to take the training again, whenever. | neutral | neutral |
| | B | Eh, do you believe that? | surprise | surprise |
| | A | Yeah? | neutral | neutral |

Table 3: Case studies from MELD test dataset on CoMPM. Red refers to the utterances of the mispredicted emotions. Blue indicates an utterance that has different emotions for the same utterance in different conversation sessions.

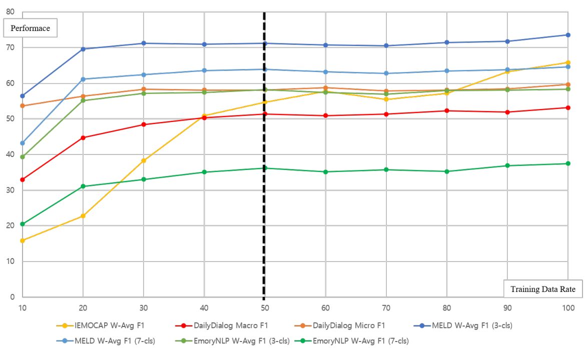


Figure 3: Performance of CoMPM according to the size of training data

| Models | IEMOCAP | EmoryNLP | |
|---|---|---|---|
| | W-Avg F1 | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) |
| CoM(sm)PM | 61.19 (-4.6) | 54.83 (-3.52) | 32.69 (-4.75) |
| CoM(la)PM | - | 57.68 (-0.67) | 35.91 (-1.53) |

Table 4: CoM(sm)PM and CoM(la)PM are the backbones of CoM as GPT2-small and GPT2-large, respectively. We were not able to experiment with CoM(la)PM in IEMOCAP due to the lack of GPU memory. The value in parentheses is the difference in performance from the original CoMPM.

### 4.5 Case Study

Table 3 illustrates the case study of the MELD. In Conv#1, CoMPM predicts the emotion of A's first utterance, *"Eh..., I don't, I don't know."*, as *neutral*, but the actual emotion is *sadness*. In this case, it is difficult to understand the context of the conversation because there are no previous utterances. So we can consider it probable the emotion is predicted as *neutral* instead of *sadness*. However, since the MELD dataset is built based on multi-modal, it is labeled by considering not only the text but also the visual information as well. In other words, studies focusing on text-level emotion recognition may suffer from such limitations, and we think that these cases can be improved in emotion recognition considering multi-modal information.

### 4.6 Training with Less Data

Recent studies improve performance by leveraging external structured knowledge, but these external sources have limitations that are mostly provided only in English. CoMPM is an approach that

| Transfer dataset | IEMOCAP | EmoryNLP | |
|---|---|---|---|
| | W-Avg F1 | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) |
| All ERC dataset | 67.47 (+1.68) | 58.76 (+0.41) | 38.02 (+0.58) |

Table 5: CoMPM is first pre-trained on all datasets and then fine-tuned in IEMOCAP and EmoryNLP.

eliminates dependence on external sources and is easily extensible to any language. However, the insufficient number of emotional data available in other countries remains a problem. Therefore, we confirm that CoMPM is effective even when the number of data is small. Figure 3 shows the performance of the model according to the number of training data and shows good performance even when only 50% of training data is used in the dataset except for IEMOCAP. IEMOCAP has a sensitive result to the ratio of training data because the total number of training data is too small.

### 4.7 Change of Backbone in CoM

We experimented by changing CoM's backbone to another pre-trained language model. Table 4 shows a comparison of GPT2-(small, medium, large) in the EmoryNLP and IEMOCAP datasets where the number of dialogues is relatively small.

We infer that GPT2-small has a lower ability to extract generalized representations than GPT2-medium, so emotional recognition performance is degraded. GPT2-large has more parameters than GPT2-medium, so its ability to extract representation is generally good, but CoM(la)CK has poorer performance than CoMPM. We infer this reason as having too many parameters for the amount of training data. In fact, there is no significant difference in sentiment classification performance in EmoryNLP, which has more training data per class. Recently, NLP researchers have been increasingly interested in pre-trained language models and have done a lot of research. We can also expect a higher performance by using a more appropriate language model for ERC datasets.

### 4.8 Transfer Learning

This section introduces a transfer learning experiment on the IEMOCAP and EmoryNLP datasets with a small number of learnable dialogues as in Section 4.7. Table 2 shows that performance is improved when external knowledge is leveraged through ATOMIC, NRC_VAD, and pre-trained language models. Therefore, we try to improve performance in a scenario where data is limited by using other emotion recognition data as external data. We first pre-train CoMPM with randomly shuffled data by summing all ERC data. Then, CoMPM is fine-tuned for each data and Table 5 shows improved performance.

The number of classes between IEMOCAP and other ERC datasets (MELD, DailyDialog, EmoryNLP) is different, and taxonomies are different even though the number of classes is the same. Therefore, only the matrix $\mathbf{M}$ of Equation 8 is newly initialized and the remaining parts are trained by transfer learning. Training is done in the same experimental environment as the original CoMPM, and the model converges quickly. As a result, the performance of CoMPM is improved by +1.68 and (+0.41, +0.58) in IEMOCAP and EmoryNLP, respectively.

### 4.9 ERC in other languages

Previous studies mostly utilize external knowledge to improve performance, but these approaches require additional publicly available data, which are mainly available for English. Indeed, structured knowledge and ERC data are lacking in other languages. Our approach can be extended to other languages without building additional external knowledge, and achieves better performance than simply using a pre-trained model. Details are in the Appendix A.

## 5 Conclusion

We propose CoMPM that leverages pre-trained memory using a pre-trained language model. CoMPM consists of a context embedding module (CoM) and a pre-trained memory module (PM), and the experimental results show that each module is effective in improving the model performance. CoMPM outperforms baselines and achieves competitive performance in all dyadic- and multi-party datasets. We compare the two methods of combining CoM and PM, and find out that close utterances of speakers in dialogues are more important for emotion recognition. In addition, we confirm that the possibility of performance improvement remains through experiments of other CoM's backbone and transfer learning.

In addition, our approach is an effective method that can be used not only in English, but also in various languages. Our approach shows competitive performance even without insufficient data or structured knowledge for actual service.

## References


Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2020. EDA: Enriching emotional dialogue acts using an ensemble of neural annotators. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 620–627, Marseille, France. European Language Resources Association.

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. *Lang. Resour. Evaluation*, 42(4):335–359.

D Datcu and LJM Rothkrantz. 2014. *Semantic audiovisual data fusion for automatic emotion recognition*, pages 411–435. Blackwell, United Kingdom.

P. Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

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.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

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

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 8002–8009. AAAI Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

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

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13622–13623.

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*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Wentao Ma, Yiming Cui, Nan Shao, Su He, Wei-Nan Zhang, Ting Liu, Shijin Wang, and Guoping Hu. 2019. TripleNet: Triple attention network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 737–746, Hong Kong, China. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik

Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. 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, Hong Kong, China. Association for Computational Linguistics.

S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

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.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8665–8672.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *arXiv preprint arXiv:1906.08487*.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. 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 *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, volume WS-18 of *AAAI Workshops*, pages 44–52. AAAI Press.

Rohola Zandie and Mohammad H. Mahoor. 2020. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. In *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference, Originally to be held in North Miami Beach, Florida, USA, May 17-20, 2020*, pages 276–281. AAAI Press.

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.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018a. Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4595–4601. International Joint Conferences on Artificial Intelligence Organization.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Kong, China. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

# A ERC in Korean Dataset

## A.1 Dataset

We constructed data composed of two speakers in Korean, and emotion-inventory is given as "surprise, fear, ambiguous, sad, disgust, joy, bored, embarrassed, neutral". The total number of sessions is 1000, and the average number of utterance turns is 13.4. We use the data randomly divided into train:dev:test in a ratio of 8:1:1. This dataset is for actual service and is not released to the public.

## A.2 Results

| Models | Korean |
|---|---|
| | W-Avg F1 |
| PM | 31.86 |
| CoM | 57.46 |
| CoMPM | 60.66 |

Table 6: Results of our approaches in Korean.

In Korean, our results are shown in Table. 6. The backbone of PM and the backbone of CoM are korean-BERT and korean-GPT owned by the company, respectively. In the Korean dataset, like the English dataset, the performance is good in the order of CoMPM, CoM, and PM. PM and CoM are not much different from fine-tuned pre-trained model. CoMPM treats the PM as a memory and predicts the final emotion by tracking the speaker's emotional state. Our approach can significantly improve baselines, and works well in other languages as well as English data.