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

Joosung Lee, Wooin Lee
Kakao Enterprise Corp., South Korea
{rung.joo, dan.kes}@kakaoenterprise.com

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. Many recent approaches show performance improvement by combining knowledge into modules learned from external structured data. However, structured data is difficult to access in non-English languages, making it difficult to extend to other languages. Therefore, we extract the pre-trained memory using the pre-trained language model as an extractor of external knowledge. We introduce CoMPM, which combines the speaker's pre-trained memory with the context model, and find that the pre-trained memory significantly improves the performance of the context model. CoMPM achieves the first or second performance on all data and is state-of-the-art among systems that do not leverage structured data. In addition, our method shows that it can be extended to other languages because structured knowledge is not required, unlike previous methods. Our code is available on github [1].

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.

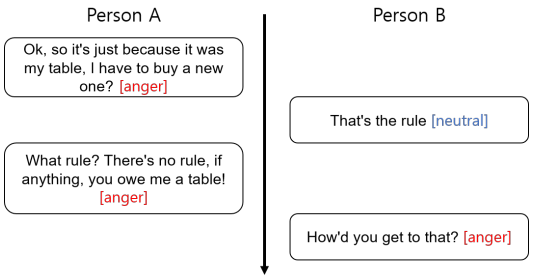

Figure 1: An example of MELD dataset

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 does 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.

Many recent studies use external knowledge to improve the ERC performance. However, this exter-

[1] https://github.com/rungjoo/CoMPM

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5669 - 5679
July 10-15, 2022 ©2022 Association for Computational Linguistics

nal 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. Since pre-trained language models are trained by unsupervised learning, these models are relatively usable approaches regardless of language types. Petroni et al. (2019) introduces that these language models can be used as knowledge bases and have many advantages over the structured knowledge bases. Based on these studies, we eliminate the dependence on structured external data used in cutting-edge systems and use a pre-trained language model as a feature extractor of knowledge.

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 perform 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 and show the performance of CoMPM when the number of data is limited.

2 Related Work

Many recent studies use external knowledge to improve the ERC performance. KET (Zhong et al., 2019) 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. COSMIC (Ghosal et al., 2020) and Psychological (Li et al., 2021) improve 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. ToDKAT (Zhu et al., 2021) improves performance by combining commonsense knowledge using COMET and topic discovery using VHRED (Serban et al., 2017) to the model.

Ekman (Ekman, 1992) constructs taxonomy of six common emotions (Joy, Sadness, Fear, Anger, Surprise, and Disgust) from human facial expressions. In addition, Ekman explains that a multimodal 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. Multimodal 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 is 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's knowledge.

If the CoM and PM are based on different backbones, we consider them to be unaligned with respect to each other's output representations. Therefore, we design the PM to follow CoM so that the output representations of CoM and PM can mutually understand each other. If CoM and PM are based on different architectures, CoMPM is trained to understand each other's representations by matching dimensions using $\mathbf{W_p}$ in Equation 4. The combination of CoM and PM is described in Section 4.5.

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.

We use an Transformer encoder as a context model. 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 RoBERTa (Liu et al., 2019). RoBERTa 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 beginning of the input and the output of the context model is as follows:

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

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 CoM.

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. Inspired by previous studies that the speaker's knowledge helps to judge emotions, we extract and track pre-trained memory from the speaker's previous utterances to utilize the emotions 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.

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:

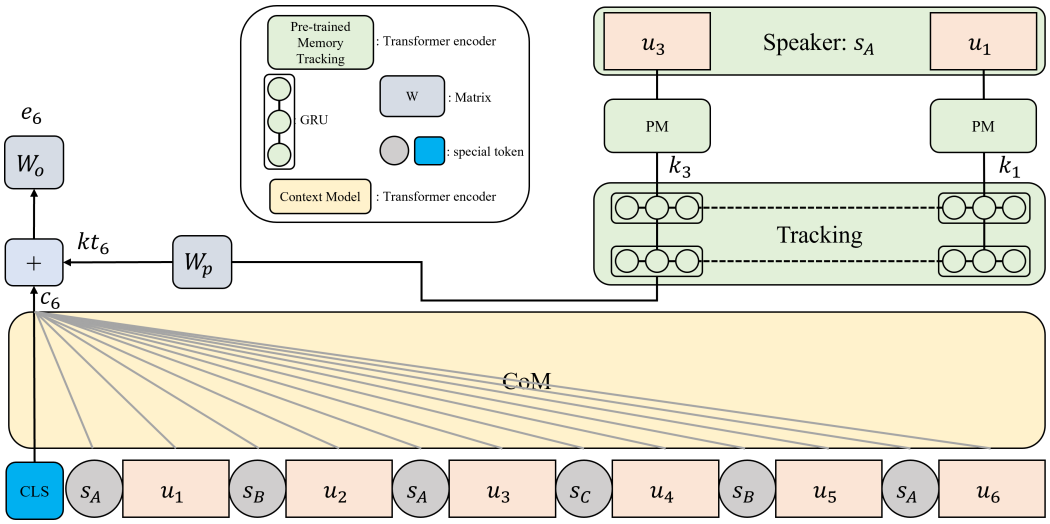

Figure 2: Our model consists of two modules: a context embedding module and a pre-trained memory module. The figure shows an example of predicting emotion of u_6, from a 6-turn dialogue context. A, B, and C refer to the participants in the conversation, where $s_A = p_{u_1} = p_{u_3} = p_{u_6}, s_B =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

$$\boldsymbol{k}_i = \text{PM}(< \text{cls} >, u_i) \tag{2}$$

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 PM.

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 recognition.

3.5.1 Tracking Method

We use $\boldsymbol{k}_i$ 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:

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

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.

$$\boldsymbol{o}_t = \boldsymbol{c}_t + \mathbf{W_p}(\boldsymbol{kt}_t) \tag{4}$$

where, $\mathbf{W_p}$ is a matrix that projects the pre-trained memory to the dimension of the context output, and is used only when PM and CoM are different pre-trained language models.

3.5.2 Emotion Prediction

Softmax is applied to the vector multiplied by $\boldsymbol{o}_t$ and the linear matrix $\mathbf{W_o} \in \mathbb{R}^{h_e \times h_c}$ 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.

$$P(e) = \text{softmax}(\mathbf{W_o}(\boldsymbol{o}_t)) \tag{5}$$

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

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.

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

We use the pre-trained model from the huggingface library [2]. 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.

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.

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.

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.

RGAT+P (Ishiwatari et al., 2020) (relational graph attention networks) proposes relational position encodings with sequential information reflecting the relational graph structure, which shows that both the speaker dependency and the sequential information can be captured.

HiTrans (Li et al., 2020) proposes a transformer-based context- and speaker-sensitive model. Hi-

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

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)
KET*	59.56	-	53.37	-	58.18	-	34.39
RoBERTa DialougeRNN	64.76	49.65	57.32	72.14	63.61	55.36	37.44
RGAT+P	65.22	-	54.31	-	60.91	-	34.42
HiTrans	64.5	-	-	-	61.94	-	36.75
COSMIC*	65.28	51.05	58.48	**73.2**	65.21	**56.51**	38.11
ERMC-DisGCN	-	-	-	-	64.22	-	36.38
Psychological*	**66.96**	51.95	**59.75**	-	65.18	-	38.8
DialogueCRN	66.05	-	-	-	58.39	-	-
ToDKAT*	61.33	**52.56**	58.47	-	**65.47**	-	**43.12**
CoMPM	66.33	**53.15**	**60.34**	**73.08**	**66.52**	**57.14**	37.37
CoM	65.05	51.17	58.63	71.67	64.9	56.27	36.34
PM	52.56	49.08	56.23	69.21	63.4	53.87	35.48
CoMPM(f)	**69.46**	51.67	59.02	73.04	65.77	55.44	**38.93**
CoMPM(s)	64.68	48.86	55.81	71.97	65.12	53.66	34.72
CoMPM(k)	64.3	52.33	59.09	72.67	66.22	56.62	36.96

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. Bold text indicates the best performance in each part (comparative models or ours). * indicates models that leverages structured external data.

Trans utilize BERT as the low-level transformer to generate local utterance representations, and feed them into another high-level transformer.

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.

ERMC-DisGCN (Sun et al., 2021) proposes a discourse-aware graph neural network that utilizes self-speaker dependency of interlocutors as a relational convolution and informative cues of dependent utterances as a gated convolution.

Psychological (Li et al., 2021) uses commonsense knowledge as enrich edges and processes it with graph transformer. Psychological performs emotion recognition by utilizing intention of utterances from not only past contexts but also future context.

DialogueCRN (Hu et al., 2021) introduces an intuitive retrieving process, the reasoning module, which understands both situation-level and speaker-level contexts.

ToDKAT (Zhu et al., 2021) proposes a language model with topic detection added, and improves performance by combining it with commonsense knowledge. The performance of ToDKAT in MELD was re-released on github [3].

[3]https://github.com/something678/TodKat

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 considering only the dialogue context. PM used alone is not used as a memory module, but the same backbone is used. PM used alone predicts emotion only with the utterance of the current turn without considering the context. CoMPM is a model in which both CoM and PM parameters are updated in the initial state of the pre-trained LM. CoMPM(f) is a model in which PM parameters are frozen in the initial state (pre-trained LM) and is not trained further, and CoMPM(s) is a model in which PM is trained from scratch. CoMPM(k) is a model in which PM is trained on ConceptNet. Following previous studies, we use the average vector for each token in PM(k) as the feature of the utterance. We use the pre-trained model provided by the site [4] as PM(k).

The effect of PM can be confirmed through the performance comparison between CoM and CoMPM, and the effect of CoM can be confirmed by comparing the results of CoM and PM. Since PM does not consider context, it showed worse performance than CoM, and the performance gap is larger in the IEMOCAP dataset with a higher average number of conversation turns. As a result, we show that the combination of CoM and PM is effective in achieving better performance.

We confirm the effect of PM structure in the model through the performance of CoMPM(s).

[4]https://huggingface.co/HungChau/distilbert-base-uncased-concept-extraction-kp20k-v1.2-concept-extraction-allwikipedia-v1.0

If PM parameters are not frozen and are instead randomly initialized (i.e. PM(s)), the performance deteriorates. CoMPM(s) performs worse than CoMPM, and even performs worse than CoM on the other datasets except for MELD. That is, PM(s) cannot be regarded as a pre-trained memory because the parameters are randomly initialized, and simply increasing the model complexity does not help to improve the performance. CoMPM(f) shows similar performance to CoMPM and achieves better performance depending on the data. PM(f) is not fine-tuned on the data, but it extracts general pre-trained memory from a pre-trained language model. The comparison between PM and PM(f) will be further described in Section 4.6. In addition, CoMPM(k) shows better performance than CoM, PM, and CoMPM(s) except for IEMOCAP. In IEMOCAP, CoMPM(k) has lower performance than CoM. For all datasets, CoMPM(k) performs slightly worse than CoMPM. In other words, ConceptNet improves the performance of CoMPM, but is not as effective as pre-trained memory. As a result, we regard pre-trained memory as compressed knowledge, which can play a role similar to external knowledge used in cutting-edge systems.

The best performance of our approaches is CoMPM or CoMPM(f), both of which combine pre-trained memory. We achieve state-of-the-art performance among all systems that do not leverage structured external data and achieve the first or second performance even including systems that leverage external data. Therefore, our approach can be extended to other languages without structured external data as well, which is described in Section 4.7.

4.5 Combinations of CoM and PM

We experiment with the effect of pre-trained memory of different language models. To eliminate the influence of the PM structure, we freeze the parameters of PM and use it as a feature extractor. Table 3 shows the performance of the pre-trained memory extracted by the different language models. If PM and CoM are based on different backbones, the pre-trained memory is projected through $\mathbf{W_p}$ as the dimension of the context output. RoBERTa+BERT and RoBERTa+GPT2 (combination of CoM and PM(f)) have lower performance than RoBERTa+RoBERTa, which is inferred because pre-trained memory of RoBERTa contains

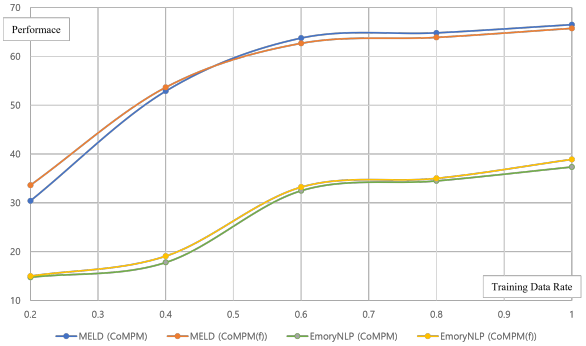

Figure 3: Performance according to the size of training data of MELD and EmoryNLP

richer information than BERT and GPT2. Since there is a lot of training data in the diallydialog and $\mathbf{W_p}$ is fine-tuned to the data to mutually understand the pre-trained memory and context representation. Therefore, we infer that performance does not decrease even if the PM changes from the dailydialog. However, even if other PMs are used, the performance is improved compared to using only CoM, so the pre-trained memory of other language models is also effective for emotion recognition.

BERT+RoBERTa has a larger performance decrease than RoBERTa+BERT. In particular, in IEMOCAP data with a long average number of turns in the context, the performance deteriorates significantly. In addition, the performance of BERT+RoBERTa is lower than CoM (RoBERTa), which supports that the performance of CoM is a more important factor than the use of pre-trained memory. In other words, we confirm that CoM is more important than PM in our system for performance, and it is effective to focus on context modeling rather than external knowledge in the study of emotion recognition in conversation.

4.6 Training with Less Data

CoMPM is an approach that eliminates dependence on external sources and is easily extensible to any language. However, the insufficient number of emotional data available in other countries (or actual service) remains a problem. Therefore, we conduct additional experiments according to the use ratio of training data in MELD and EmoryNLP, where there is neither too much nor too little data. Figure 3 shows the performance of the model according to the ratio of the training data. In MELD and EmoryNLP, even if only 60% and 80% are used, respectively, the performance decreases by only 3 points.

CoM	PM(f)	IEMOCAP	DailyDialog		MELD	EmoryNLP
		W-Avg F1	Macro F1	Micro F1	W-Avg F1 (7-cls)	W-Avg F1 (7-cls)
RoBERTa	BERT	65.93 (-3.53)	52.74 (+1.07)	59.97 (+0.95)	65.41 (-0.36)	37.25 (-1.68)
RoBERTa	GPT2	68.54 (-0.92)	50.68 (-0.99)	59.61 (+0.59)	65.58 (-0.19)	36.39 (-2.54)
BERT	RoBERTa	62.69 (-6.77)	48.99 (-2.68)	57.34 (-1.68)	63.79 (-1.98)	35.47 (-3.46)

Table 3: Emotion recognition performance according to the combination of different backbones of CoM and PM. The value in parentheses is the performance difference from the original CoMPM(f) (RoBERTa + RoBERTa). We use the bert-large-uncased and GPT2-medium versions.

Table 2 shows that CoMPM(f) achieves better performance than CoMPM in the emotion classification of IMEOCAP and EmoryNLP, which has fewer training data than other settings. On the other hand, if there is a lot of training data, CoMPM shows better performance. Figure 3 shows that as the number of data decreases, CoMPM(f) shows better results than CoMPM, which indicates that it is better to freeze the parameters of PM when the number of training data is insufficient. Therefore, if there is a lot of training data in the real-world application, CoMPM is expected to achieve good performance, otherwise it is CoMPM(f).

4.7 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.

4.7.1 Korean 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.

4.7.2 Results in the Korean Dataset

In Korean, our results are shown in Table 4. The backbone of CoM and PM is Korean-BERT 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. Our approach simply shows a significant performance improvement on baselines that are fine-tuned to the language model and works well for other languages as well as for English.

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

Table 4: Results of our approaches in Korean.

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 performance. CoMPM outperforms baselines on both dyadic-party and multi-party datasets and achieves state-of-the-art among systems that do not use external knowledge. In addition, CoMPM achieves performance comparable to cutting-edge systems that leverage structured external knowledge, which is the effect of pre-trained memory of the language model.

By combining other pre-trained memories, we find that the pre-trained memory extracted with RoBERTa is richer and more effective than the pre-trained memory extracted with BERT or GPT2. Since we believe that pre-trained memory is proportional to the performance of a language model, a language model with a large training corpus and many parameters is considered to be more effective. However, we find that context modeling is more important than pre-trained memory for emotion recognition in conversation, and future research will focus on context modeling.

Additionally, our approach achieves competitive

performance and does not require externally structured data. Therefore, we show that it can be easily extended to Korean as well as English, and it is expected to be effective in other countries.

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.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, Online. 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.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200, Barcelona, Spain (Online). International Committee on 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.

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.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

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.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.