Contextual Augmentation of Pretrained Language Models for Emotion Recognition in Conversations

Jonggu Kim **Hyeonmok Ko** **Seoha Song** **Saebom Jang** **Jiyeon Hong**
Samsung Research
Seoul, Republic of Korea
{jonggu88.kim,felix.ko,seoha.song,saebom.jang,ji-yeon.hong}@samsung.com

Abstract

Since language model pretraining to learn contextualized word representations has been proposed, pretrained language models have made success in many natural language processing tasks. That is because it is helpful to use individual contextualized representations of self-attention layers as to initialize parameters for downstream tasks. Yet, unfortunately, use of pretrained language models for emotion recognition in conversations has not been studied enough. We firstly use ELECTRA which is a state-of-the-art pretrained language model and validate the performance on emotion recognition in conversations. Furthermore, we propose contextual augmentation of pretrained language models for emotion recognition in conversations, which is to consider not only previous utterances, but also conversation-related information such as speakers, speech acts and topics. We classify information based on what the information is related to, and propose position of words corresponding to the information in the entire input sequence. To validate the proposed method, we conduct experiments on the DailyDialog dataset which contains abundant annotated information of conversations. The experiments show that the proposed method achieves state-of-the-art F1 scores on the dataset and significantly improves the performance.

1 Introduction

As voice assistants are widely used, emotion recognition is also emerging as an important technique to provide a rich user experience. Considering that it can detect the emotion state of speakers in real-time in an on-going conversation, it can be utilized in a variety of applications to generate more diverse and appropriate responses. As if to reflect this proliferation, detecting emotional state and emotion change in conversations has been widely studied. However, there is no significant progress in this research area. Previous studies focus extensively on emotion detecting using groups of words and/or utterances representing emotion, but those are not sufficient to detect the emotion state change that varies depending on dialogue subject or speaker in given conversations.

The proposed system in this paper models each conversation to contain two different levels of information (Fig. 1). Conversation-level information is effective in distinguishing conversations that show different aspect of emotion changes. Given the specific type of conversation, utterance-level information is used to recognize speaker's emotion. Basically, our research is based on the hypothesis that even if it's the same conversation there will be very different results depending on what previous or current conversations relate to and/or who is the speaker now. They so far have not considered essential information enough, but our results show the model trained with the essential information outperforms the state-of-the-art.

Various networks that could capture features that fit our hypothesis were considered, our evaluation results show that the Transformer (Vaswani et al., 2017) encoder family, which consists of multi-layers of self-attention and fully-connected layer, generally shows better performance. A self-attention model learns to generate representations for each token based on the context of the token. That is, the information mentioned above can be incorporated as each token representing fragments of information.

Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media, pages 64–73
Barcelona, Spain (Online), December 13, 2020.

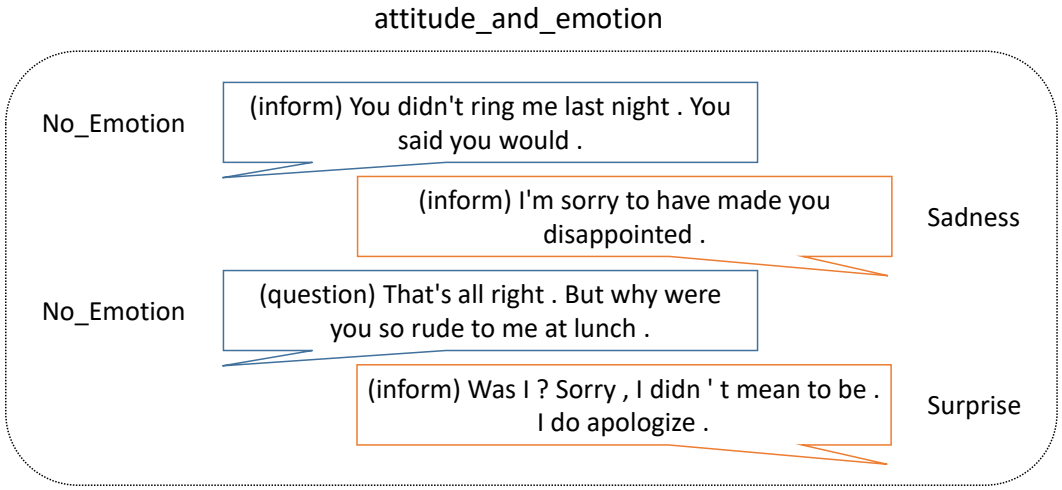

Figure 1: An example of conversation-level information and utterance-level information in DailyDialog. Topics like "attitude_and_emotion" belong to the conversation-level information, and speech acts like "inform" and "question" belong to the utterance-level information. Speakers also belong to the utterance-level information, which are regularly changed.

In this paper, we use ELECTRA (Clark et al., 2020) which is a state-of-the-art pretrained language model based on the Transformer encoder. Our experiments with the DailyDialog dataset (Li et al., 2017) containing enough annotated information of English conversations show the proposed model significantly improves the F1 scores comparing to the previous studies. Also, we thoroughly analyze effectiveness of incorporating each type of information as words in ELECTRA and different context lengths and present the result.

2 Related Work

Convolutional neural networks (CNN) and recurrent neural networks (RNN) have traditionally been used for emotion recognition. DialogueRNN (Majumder et al., 2019) has been proposed to model a speaker, a context from the preceding utterances and an emotion of the preceding utterances. DialogueGCN (Ghosal et al., 2019) has been proposed to model a conversation using directed graph to propagate the speaker's dependency information. Contextual information among distant utterances is propagated through two consecutive convolution operations by providing a convolutional network with the graph (Defferrard et al., 2016). Attention gated hierarchical memory network (AGHMN) (Jiao et al., 2020) has been proposed for real-time emotion recognition with an hierarchical memory network (HMN), a bidirectional gated recurrent unit (BiGRU) as the utterance reader and a BiGRU fusion layer for interaction between historical utterances; this network includes an attention gated recurrent unit (GRU) to update internal state and a bidirectional variant GRU to keep a balance between the contextual information from recent memories and that from distant memories. Recently, a generalized neural tensor block followed by a two-channel classifier is designed to perform contextual compositionality which obtains context information and incorporates the context into utterance representation and sentiment classification simultaneously (Li et al., 2020b).

Many natural language processing (NLP) tasks have applied Transformer (Vaswani et al., 2017) to capture long context information and enrich text representations. Knowledge-enriched Transformer (KET) (Zhong et al., 2019) uses hierarchical self-attention for exploiting contextual information and dynamically refers to external commonsense knowledge. HiTransformer-s (Li et al., 2020a) uses a hierarchical Transformer network with speaker embeddings to capture the contextual information and the interaction of speakers.

There are many studies on how to use various data such as visual expression, voice, and text to improve the performance of emotion recognition. Each type of data can be used alone in the emotion classification

task, and when used together, better performance is observed empirically. This multimodal technique simply uses text and voice related to the speech (Ho et al., 2020), or additionally uses visual information such as the speaker's expression that occurs at the same time as the speech (Zadeh et al., 2018; Mittal et al., 2020; Delbrouck et al., 2020). Models that consider multimodality combine and use data of different properties for one target task. Among the studies for emotion recognition, there are studies that perform classification by hierarchically combining the relations between modals (Zadeh et al., 2018), or to select valid features using relations between modals (Mittal et al., 2020). Attention mechanisms, which are showing good performance in recent years, have also been used in several studies to find the relationship between modals (Ho et al., 2020; Delbrouck et al., 2020).

In sum, previous studies also use pretrained models and do not only concentrate on improving performance with neural networks, but incorporates a large deal of varying contextual information. Likewise, we propose and validate a method to improve performance in this paper. However, compared to the previous studies, our proposed approach has a relative strength: it is easy to incorporate the approach to finetune state-of-the-art pretrained language models.

3 Background

Pretrained language models like BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020) use the encoder of Transformer (Vaswani et al., 2017) that consists of multi-layers of self-attention and fully-connected layer. Attention can be defined using three terms, Q (queries), K (keys) and V (values). Self-attention is an attention method where Q, K and V are the same or generated from the same source. Given a sequence of vectors $[y_1, y_2, ..., y_{N_{token}}]$ whose length N_{token} is the same.

Performing projections of the queries, keys and values respectively to d_q, d_k and d_v dimensions N_{head} times[1] is empirically more effective than a single linear projection, and this method is called multi-head attention. Because the given sequence of vectors can be packed into a single matrix $X \in \mathbb{R}^{N_{token} \times d_x}$, the produced sequence of vectors can also be represented as a matrix $Y \in \mathbb{R}^{N_{token} \times d_y}$. Given X, Y is computed in the multi-head attention way as:

$$Q_i = X^T W_i^Q, \tag{1}$$
$$K_i = X^T W_i^K, \tag{2}$$
$$V_i = X^T W_i^V, \tag{3}$$
$$head_i = \mathrm{Softmax}(Q_i K_i^T) V_i, \tag{4}$$
$$Y = \mathrm{Concat}(head_1, ..., head_{N_{head}}) W^O, \tag{5}$$

where $[\cdot]^T$ is transpose of $[\cdot]$, $W_i^Q \in \mathbb{R}^{N_{token} \times d_q}$, $W_i^K \in \mathbb{R}^{N_{token} \times d_k}$, $W_i^V \in \mathbb{R}^{N_{token} \times d_v}$ and $W^O \in \mathbb{R}^{N_{head} \times d_{model}}$ are trainable weight matrices.

In each layer, a fully-connected layer is used to linearly transform the output of the self-attention layer, and layer normalization is used after the two outputs of self-attention and fully-connected layer are added. By stacking the layer L times, the entire model is built.

By (pre)training on a large amount of text data, a self-attention model learns to generate representations for each token based on the context of the token. Then the pretrained model learns to generate task-specific answers by finetuning. For example, by finetuning, ELECTRA can learn to recognize the emotion of the given utterance or to classify the polarity of individual words in the utterance.

ELECTRA that we used for the experiments is different from other pretrained language models in that it uses not the generator, but an additional model, a discriminator, which consists of the encoder of Transformer as the same described above. The difference between them is about details for pretraining, not finetuning, so we omit the explanation. Details of the model related to our proposed method is introduced in the next section.

[1] $d_k = d_v = d_{model}/N_{head}$ is generally used.

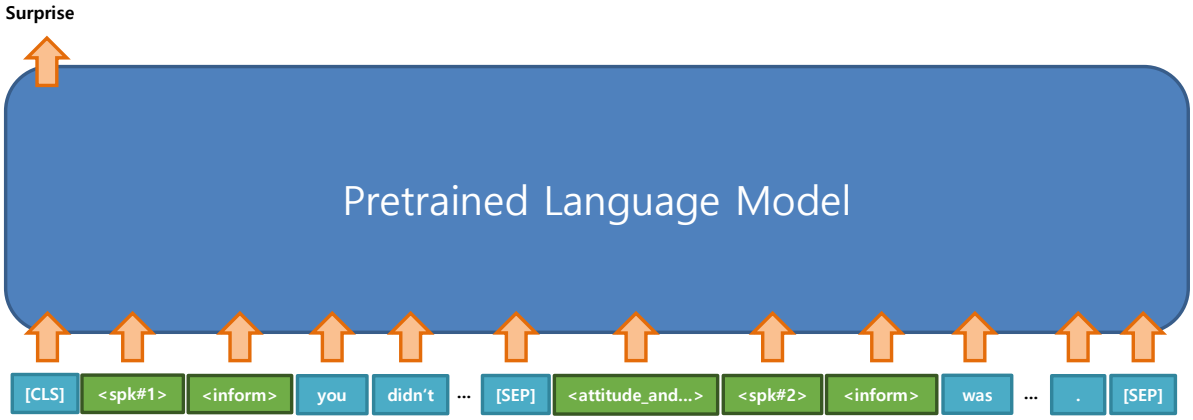

Figure 2: An example of input/output for contextual augmentation of pretrained language models.

4 Proposed Method

By finetuning on a downstream task, pretrained language models are optimized for the task. Two types of input form are generally used, and which type is used depends on the downstream task. The two types are "`[CLS] seg_A [SEP]`" and "`[CLS] seg_A [SEP] seg_B [SEP]`", where `[CLS]` is a special symbol and `[SEP]` is a special separator token.

In this paper, emotion recognition in conversations is considered, which is a task that when given T-th utterance and the context (previous utterances), the model recognizes the emotion implicit in the T-th utterance. For this task, we segment all utterances into two parts, a part for the T-th utterance and a part for the context.

Specifically, the T-th utterance is decomposed into a sequence of N_T words $[w_1^T, w_2^T, ..., w_{N_T}^T]$, where $N_{[.]}$ is the number of words of the [.]-th utterance. We put them all into `seg_B`. On the other hand, words of the previous utterances $[w_1^{T-M}, w_2^{T-M}, ..., w_{N_{T-1}}^{T-1}]$ are put into `seg_A`, where M is the context length. In sum, the input form is "`[CLS]` w_1^{T-M} w_2^{T-M} ... $w_{N_{T-1}}^{T-1}$ `[SEP]` w_1^T w_2^T ... $w_{N_T}^T$ `[SEP]`".

Each type of information is converted into words, and then put into appropriate positions of the information. Then the model will consider all the information together.

We can think of two levels of information, utterance-level and conversation-level. In DailyDialog (Li et al., 2017), for example, there are speakers, speech acts and topics, so they can be respectively mapped to one of the information levels: speakers and speech acts are mapped to utterance-level information, and topics are mapped to conversation-level information.

We add $\texttt{UTT}_{TYPE}$ and $\texttt{CON}_{TYPE}$ to the input sequence, encoding utterance-level and conversation-level information. For utterance-level information, we add a symbol of the information of the t-th utterance $\texttt{UTT}_{TYPE}^t$ in front of the utterance. Thus, we add a symbol of the speaker of the t-th utterance $\texttt{UTT}_{spk}^t$ in front of it for the speaker information. In the same way, we add a symbol of the speech act of the t-th utterance $\texttt{UTT}_{act}^t$ in front of the utterance for the speech act information.

For conversation-level information, we add a symbol of the information $\texttt{CON}_{TYPE}$ in front of `seg_B`. For example, to incorporate a topic of the conversation, we add a symbol of the topic $\texttt{CON}_{topic}$ in front of the T-th utterance.

As a result, the full input form for DailyDialog is "`[CLS]` $\texttt{UTT}_{spk}^{T-M}$ $\texttt{UTT}_{act}^{T-M}$ w_1^{T-M} w_2^{T-M} ... $w_{N_{T-M}}^{T-M}$ $\texttt{UTT}_{spk}^{T-1}$ $\texttt{UTT}_{act}^{T-1}$ w_1^{T-1} w_2^{T-1} ... $w_{N_{T-1}}^{T-1}$ `[SEP]` $\texttt{CON}_{topic}$ $\texttt{UTT}_{spk}^{T}$ $\texttt{UTT}_{act}^{T}$ w_1^T w_2^T ... $w_{N_T}^T$ `[SEP]`". The model then generates an emotion label at the position of `[CLS]` at the last layer (Fig. 2). As the input is fed to ELECTRA, we expect the model to consider and relate the information to all tokens in the utterance via multi-layers of self-attention for emotion recognition.

Table 1: The number of utterances for emotion types in DailyDialog.

	Emotion						
Split	**Anger**	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**	**No_Emotion**
train	827	303	146	11,182	969	1,600	72,143
valid	77	3	11	684	79	107	7,108
test	118	47	17	1,019	102	116	6,321
total	1,022	353	174	12,885	1,150	1,823	85,572

Table 2: The number of utterances for speech act types in DailyDialog.

Speech Act	**train**	**valid**	**test**	**total**
Inform	39,873	3,125	3,534	46,532
Question	24,974	2,244	2,210	29,428
Directive	14,242	1,775	1,278	17,295
Commissive	8,081	925	718	9,724

5 Experiments

5.1 Settings

We use the DailyDialog dataset (Li et al., 2017) for the experiments. DailyDialog is a human-written conversation dataset that reflects our daily communication way in various topics on our daily life. The dataset consists of 13,118 dialogues and the dialogues include 102,979 utterances. A dialog has a manually-labeled topic and an utterance has manually-labeled intention (speech act) and emotion information. In DailyDialog, emotions to be recognized are decomposed into seven categories (Anger, Disgust, Fear, Happiness, Sadness, Surprise and No_Emotion) (Table 1). Speech acts that the proposed model uses are decomposed into four categories (Inform, Question, Directive and Commissive) (Table 2), and topics are decomposed into 10 categories (Ordinary Life, School Life, Culture & Education, Attitude & Emotion, Relationship, Tourism, Health, Work, Politics and Finance) (Table 3). For fair comparison with state of the art models (Poria et al., 2017; Majumder et al., 2019; Zhong et al., 2019; Hazarikaa et al., 2021), we use the same training dataset, the same validation dataset and the same test dataset[2]: 11,118 dialogs (87,170 utterances) for training, 1000 dialogs (8,069 utterances) for validation and 1000 dialogs (7,740 utterances) for test.

We finetune and use pretrained ELECTRA-Base as a baseline model for emotion recognition. ELECTRA-Base has 12 layers, 768 hidden dimensions and 12 heads. We use batch size of 4, learning rate of 0.0001, maximum sequence length of 100 and 512 and Adam optimizer (Kingma and Ba, 2015) with 1 epoch to finetune all models[3]. As an evaluation metric, we mainly use the micro-averaged F1 score excluding the majority class, No_Emotion, because of the imbalanced class distribution, which is the same metric as in the previous work (Zhong et al., 2019). To compare with TL-ERC (Hazarikaa et al., 2021), we use the weighted macro F1 score. We use a one tailed t-test to validate the significance of improvements. Also in the comparison, we ran each model five times, and report their average scores.

We use ground-truth labels of all information. For speaker labels, we use turn numbers to distinguish a speaker from a listener because the current speaker is not explicitly given, the speaker is changed turn by turn. Then we map the numbers to specific expressions existing in the vocabulary of ELECTRA. Specifically, if the turn number is odd, we used "[unused1]"; otherwise, we used "[unused2]". For acts and topics, we used names of the labels enclosed by "<" and ">". e.g., "<question>" and "<ordinary_life>".

We compare our model with the state-of-the-art emotion recognition models. Note that all scores of

[2]The split can be found at `https://github.com/declare-lab/conv-emotion`.

[3]Maximum sequence length of 100 was used for fair comparison with BERT reported in previous work (Zhong et al., 2019) (Table 4). In the other cases, maximum sequence length of 512 was used (Table 5 and Fig. 3).

Table 3: The number of utterances for topic types in DailyDialog.

Topic	train	valid	test	total
Ordinary Life	23,587	3,507	2,162	29,256
School Life	4,257	0	299	4,556
Culture & Education	469	0	55	524
Attitude & Emotion	3,683	40	344	4,067
Relationship	29,713	512	2,582	32,807
Tourism	6,822	1,040	642	8,504
Health	1,969	458	205	2,632
Work	11,889	1,809	1,104	14,802
Politics	1,295	156	132	1,583
Finance	3,486	547	215	4,248

the state-of-the-art models are the scores reported in previous work (Zhong et al., 2019). The models are described as follows:

CNN (Kim, 2014): A single-layer of convolutional neural networks for the current utterance. The model does not use contextual information.

CNN + cLSTM (Poria et al., 2017): A contextual LSTM (cLSTM) to capture contextual information at utterance level after a CNN layer.

BERT (Devlin et al., 2019): A Base version of BERT finetuned on emotion recognition. The difference from ELECTRA is the method of pretraining.

DialogueRNN (Majumder et al., 2019): A customized RNN model to capture speakers and context information. Several GRUs are used to track global/party state and speaker information after feature extraction from utterances by CNN. For DailyDialog, two speakers are distinguished using a turn number of each utterance.

KET (Zhong et al., 2019): A hierarchical self-attention model to encode hierarchical conversation representations. Also, the model retrieves related commonsense knowledge from external knowledge base and exploits the knowledge.

TL-ERC (Hazarikaa et al., 2021): A transfer learning-based approach. A Transformer encoder is first trained to generate multi-turn conversations, and the trained model is trained again to generate an emotion in conversations.

5.2 Results

In comparison with state-of-the-art models, ELECTRA achieves state-of-the-art F1 scores on DailyDialog (Table 4). Even if the model is similar in the model structure, obtains an F1 score higher than BERT (Devlin et al., 2019). This result shows that ELECTRA is better than BERT in emotion recognition also.

Incorporating information as words significantly improves micro/weighted F1 scores. ELECTRA with contextual augmentation achieves micro F1 of 57.97 % and weighted F1 of 55.73 % while ELECTRA achieves micro F1 of 55.13 % and weighted F1 of 51.63 %. The F1 scores of ELECTRA with contextual augmentation are the state-of-the-art F1 scores on DailyDialog.

6 Discussion

We analyze effectiveness of the proposed method in detail in this section. In the proposed model, because we use three kinds of information, speaker, act and topic, we separate them to validate effectiveness of each type of information (Table 5). In this analysis, we find which information is the most effective on DailyDialog and how different results are according to emotion categories.

Table 4: Comparison with state-of-the art models on the test dataset of DailyDialog. **: $p < 0.01$ compared to ELECTRA.

Model	micro F1	weighted F1
CNN (Kim, 2014)	49.34	-
CNN + cLSTM (Poria et al., 2017)	49.90	-
BERT (Devlin et al., 2019)	53.12	-
DialogueRNN (Majumder et al., 2019)	50.65	-
KET (Zhong et al., 2019)	53.37	-
TL-ERC (Hazarikaa et al., 2021)	-	48.00
ELECTRA	55.13	51.63
ELECTRA with Contextual Augmentation	**57.97****	**55.73****

Table 5: F1 score of incorporating speaker (S), act (A) and topic (T) as words for each emotion type on the test dataset of DailyDialog.

		F1						micro F1
#	Information Type	Anger	Disgust	Fear	Happiness	Sadness	Surprise	All
1	none	28.14	0.00	0.00	62.25	1.88	50.76	54.67
2	S	25.75	2.40	0.00	63.66	0.38	50.24	55.70
3	A	29.71	**3.87**	0.00	63.35	0.00	49.43	55.31
4	T	30.01	1.51	0.00	65.82	34.53	50.49	58.57
5	S,A	28.67	0.00	0.00	63.89	0.00	50.47	55.82
6	S,T	29.28	2.10	0.00	**66.90**	34.80	51.23	**59.20**
7	A,T	29.11	2.36	0.00	66.27	29.73	**51.28**	58.64
8	S,A,T	**29.69**	0.00	0.00	66.53	**36.15**	50.80	58.99

Additionally, we conduct experiments to evaluate performance of the model with respect to the context length that is the number of previous utterances considered. The context length is helpful information in understanding of conversations, and in guessing the emotion more correctly. We analyze how effective different context lengths are, and at the same time, how different the performance in incorporation of different information types is.

6.1 Effectiveness of Incorporating Speaker, Act and Topic

To validate effectiveness of incorporating speaker, act and topic as words, we conduct ablation study for emotion classes (Table 5). In this study, we found a few tendencies with respect to what kind of information is used.

First, all models with incorporation of topic achieve F1 scores higher than 58.0 % (row 4, 6, 7 and 8) while the others do not. This result means that incorporation of topic is effective to improve performance of emotion recognition. In analysis on F1 score for each emotion type, we found a tendency that all models with incorporation of topic achieved higher F1 scores for Happiness and Sadness than those for the other emotions. We interpret this result as a topic is implicitly related to the emotion of happiness or sadness of the speakers in the dataset. Also, we found that these gains drove micro F1 scores to being higher. All the models obtain F1 scores of 0.00 % on Fear, we ascribe this result to low distribution (146/87,170) of the emotion type in the training dataset.

Compared to the baseline method (row 1), incorporating speaker or act also obtains improvements of F1 score (row 2, 3, 5, 6, 7 and 8); however, they are not outstanding. Specifically, in the case of DailyDailog, incorporating speaker information is not helpful as it only indicates change of turn, and does not include any personalized information. This limits the improvements on this dataset, but incorporation of speaker information could help on other datasets where personalized speaker information is available.

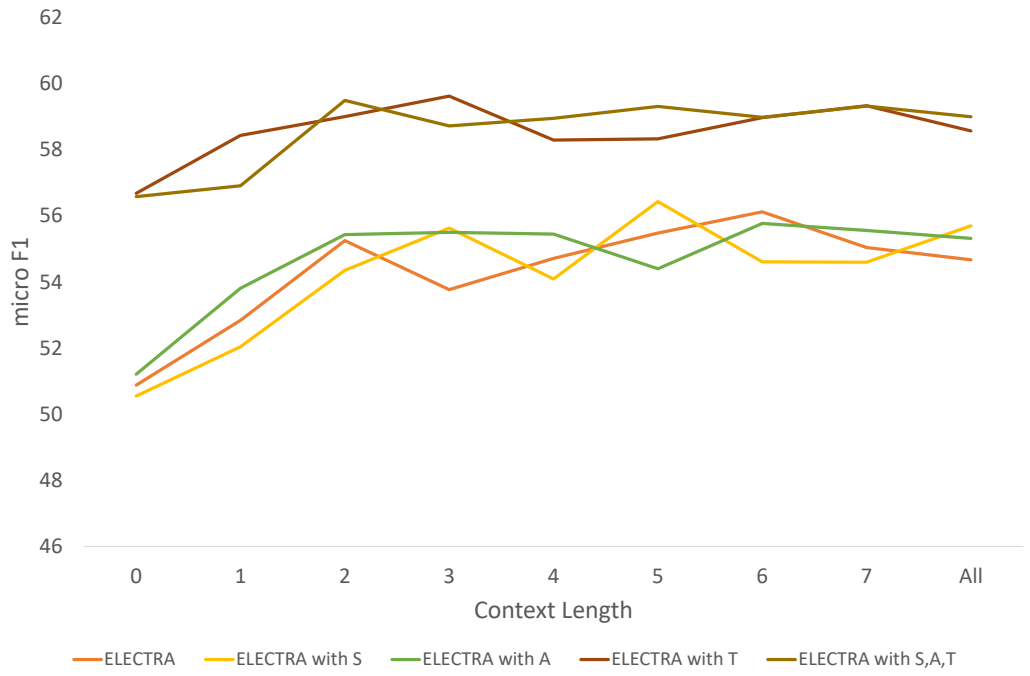

Figure 3: F1 score of incorporating speaker (S), act (A) and topic (T) for different context length on the test dataset of DailyDialog.

6.2 Effectiveness of Different Context Length with Incorporation of Different Information

What kind of help would the context length bring to emotion recognition? We analyze the effectiveness of context length with incorporation of different information (Fig. 3). Note that we use average of F1 scores after 5 runs for all context, whereas we use F1 scores after a single run for the other context lengths (0 to 7).

We found a tendency that in every model, the longer context length is used, the higher F1 score is achieved. In every model, the lowest micro F1 score is obtained using context length of 0. F1 scores are gradually improved when context length becomes longer up to 2. F1 scores of all models using context length from 2 to all are similar, but do not degrade.

7 Conclusion

We employ a state-of-the-art pretrained language model, ELECTRA, to emotion recognition in conversations. Because ELECTRA is a powerful NLP model by itself, we do not propose to modify the structure. Instead, we explore and propose an effective method of providing the model with extra information as words to improve performance of emotion recognition.

In this paper, we consider three kinds of information, speaker, act and topic, and propose a method to incorporate the information in pretrained language models. By incorporating the information, pretrained language models are expected to consider and relate the information to all tokens in the utterance via multi-layers of self-attention for emotion recognition.

Our experiments show that the proposed method improves performance of emotion recognition with large margin, and that the gain margin depends on how important the provided information is. On the DailyDialog dataset, when given the topic as words, performance of the model is highly improved. This result means that if the information provided with the proposed method is closely related to the emotion,

the performance of emotion recognition is highly improved. In other words, leveraging state-of-the-art pretrained language models by simply augmenting the features with meta-information leads to large improvements, without needing to resort to more complex modeling.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems 29*, pages 3844–3852.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

Devamanyu Hazarikaa, Soujanya Poria, Roger Zimmermanna, and Rada Mihalcea. 2021. Conversational transfer learning for emotion recognition. *Information Fusion*, 65:1 – 12.

N. Ho, H. Yang, S. Kim, and G. Lee. 2020. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, 8:61672–61686.

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-20)*.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*.

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.

Qingbiao Li, Chunhua Wu, Zhe Wang, and Kangfeng Zheng. 2020a. Hierarchical transformer network for utterance-level emotion recognition. *Applied Sciences*, 10(13):4447.

Wei Li, Wei Shao, Shaoxiong Ji, and E. Cambria. 2020b. Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis. *ArXiv*, abs/2006.00492.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 6818–6825.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 1359–1367.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

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.