EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa

Taewoon Kim and **Piek Vossen**
Vrije Universiteit Amsterdam
{t.kim,p.t.j.m.vossen}@vu.nl

Abstract

We present EmoBERTa: Speaker-Aware **Emo**tion Recognition in Conversation with Ro**BERTa**, a simple yet expressive scheme of solving the ERC (emotion recognition in conversation) task. By simply prepending speaker names to utterances and inserting separation tokens between the utterances in a dialogue, EmoBERTa can learn intra- and inter- speaker states and context to predict the emotion of a current speaker, in an end-to-end manner. Our experiments show that we reach a new state of the art on the two popular ERC datasets using a basic and straight-forward approach. We've open sourced our code and models at https://github.com/tae898/erc.

1 Introduction

The scope of emotion recognition is very wide, ranging from stills of face images, audio data to the actual utterances or text as in tweets. In this paper, we focus on emotion recognition in conversation (ERC), which is a subfield of emotion recognition. More specifically, the task is to predict the emotion of a current speaker who's engaging in a conversation with one person or more. Recognizing emotion is important to areas such as affective computing and human-robot communication, in which it can be an important feedback mechanism.

As humans use multiple sensory inputs to have a conversation (e.g., vision, voice, etc.) the ERC task can also include multiple modalities (e.g., visual, audio, text, etc.). Here, we report on our first experiments on the text modality, leaving using multiple modalities within our framework to future work.

Since the introduction of the Transformer (Vaswani et al., 2017), transformer-based deep neural network models have become the dominating neural network model in sequence modeling. Especially, pretrained encoder-only models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have shown that they can be successfully fine-tuned for downstream tasks, such as sentence classification and question answering. ERC can be seen as a special case of sequence modeling, since emotions are expected to be triggered by a preceding event in any modality.

Our approach to ERC enriches such transformer models by including the speaker identity in the sequence information spanning multiple utterances. By adapting the RoBERTa sequence representation, we improve the SOTA on two popular benchmark datasets.

2 Related Work

Most of the existing works on ERC combine different kinds of neural network architectures (e.g., CNNs, RNNs, Transformers, GNNs, etc.) (Li et al., 2020b), (Li et al., 2020c), (Ishiwatari et al., 2020), (Wang et al., 2020), (Hazarika et al., 2021), (Sheng et al., 2020), (Ghosal et al., 2020), (Ghosal et al., 2019). The biggest downside of such approaches is that each part of the model is responsible for extracting their own features. These extracted features might not be ideal for the other parts of the model. Also, since these models are combinations of sub-models, it is hard to understand what each sub-model is contributing and how to improve the overall model. Some of the approaches try to take advantage of external knowledge bases (Ghosal et al., 2020), (Zhong et al., 2019a), which adds even more complexity to the model.

A substantial number of approaches are heavily based on RNNs (e.g., GRU) to model the sequence (Jiao et al., 2019), (Lu et al., 2020). The biggest problem with this is that it inherently decouples the word embedding extraction and the sequence modeling, whereas BERT-like models tackle them at once, often leading to a better performance. Also, they have to rely on external decontextualized word embedding extractors (e.g., GloVe (Pennington et al., 2014) or word2vec (Mikolov et al., 2013)). Furthermore, training an RNN is very inefficient since “backpropagation through time” has to wait until the last input of a sequence has been processed.

The approaches that are most closely related to us are HiTrans (Li et al., 2020a) and DialogXL (Shen et al., 2020). HiTrans packs multiple utterances with `[CLS]` tokens prepended into one input sequence. This sequence is first fed into a BERT and then to another transformer. DialogXL is based on XLNet (Yang et al., 2019). Our approach differs in that we just use RoBERTa and encode the speaker information with multiple utterances.

3 Methodology

3.1 Problem Definition

Let’s say a dialogue of M utterances is given and I interlocutors are engaging in a conversation. Then the dialogue can be expressed as a list of vectors: $dialogue = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_M}]$, where an utterance $\boldsymbol{x_t}$ contains several words (tokens). Each utterance $\boldsymbol{x_t}$ is spoken by a unique $speaker \in I$. Since this is a supervised setup, every utterance $\boldsymbol{x_t}$ has one corresponding label y_t that is annotated by a human.

The simplest way to solve this problem is to come up with a function f that takes $\boldsymbol{x_t}$ as an input and outputs the correct label y_t. However, this doesn’t take context into account. It’s easily conceivable that the function f should also consider the past utterances $[\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_{t-1}}]$ or even the future utterances $[\boldsymbol{x_{t+1}}, \boldsymbol{x_{t+2}}, ..., \boldsymbol{x_M}]$ to model the context.

3.2 EmoBERTa

EmoBERTa starts from the pretrained `roberta-large` model (Liu et al., 2019). Since the task is basically a sequence classification task, we simply add a randomly initialized linear layer with the softmax nonlinearity to the first hidden state (this state corresponds to the `[CLS]` token) of the last layer of the pretrained model.

We chose RoBERTa, among the many BERT-like models, because its structure is not only relatively simple, but also it can deal with more than two segments. The original authors of RoBERTa simply used two `</s>` tokens consecutively as `[SEP]` token, which separates the first and the second segments. Although the pretrained model has not been trained on more than two segments, we show that EmoBERTa can be generalized to three segments per input sequence.

The first, second, and third segments contain the past utterances, the current utterance, and the future utterances, respectively, in a dialogue. Each utterance is prepended with the name of a speaker so that the model is aware which utterance is spoken by whom. The task is to predict the emotion of the current utterance.

RoBERTa uses `<s>` and `</s>` as `[CLS]` and `[EOS]` tokens, respectively. Building an input sequence for EmoBERTa is outlined in Algorithm 1. Example sequences can be found in Figure 1.

Algorithm 1: Building an input sequence[1]

```
Given the current utterance
  $\boldsymbol{x_t} \in \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_M}\}$;
  $max\_tokens = 512 - 2$;
$sequence =$
  $[SEP] + tokenize(\boldsymbol{x_t}) + [SEP]$;
$i = 1$;
while $len(sequence) <= max\_tokens$ do
    Prepend $speaker(\boldsymbol{x_{t-i}})$ + “ : ” + $\boldsymbol{x_{t-i}}$
      to $sequence$;
    Append $speaker(\boldsymbol{x_{t+i}})$ + “ : ” + $\boldsymbol{x_{t+i}}$
      to $sequence$;
    $i = i + 1$;
end
Remove the last appended / prepended
  utterances;
$sequence = [CLS] + sequence + [EOS]$;
```

3.3 Training

The loss is calculated as the sum of cross entropy loss and L^2 weight decay (Krogh and Hertz, 1991). We use adaptive gradient descent (Kingma and Ba, 2015), (Loshchilov and Hutter, 2019) with gradual linear warmup learning rate scheduling (Goyal et al., 2017). The peak learning rate was determined using Optuna (Akiba et al., 2019). Mixed floating point precision was used to reduce the training time and increase the batch size (Micikevicius et al., 2017). See Appendix A.1 for the details (e.g., hyperparameters, training time, hardware, etc.). We mostly used the huggingface transformer pytorch library for training (Wolf et al., 2020).

Then the training loss function is

[1] The pretrained RoBERTa model can have a maximum of 512 tokens in one input sequence. The while loop terminates if there are no more available past / future utterances to add in the dialogue. We empirically found that capitalizing the names of the interlocutors leads to slightly better results.

$$\mathcal{L}(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{C-1} y_c^{(i)} \log(\hat{y}_c^{(i)}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \tag{1}$$

where $\boldsymbol{y}^{(i)}$, is a one-hot label vector, $\hat{\boldsymbol{y}}^{(i)}$ is the softmax output vector given the input $\boldsymbol{x^{(i)}}$, N is the number of training data samples, C is the number of classes, λ is a L^2 regularization rate, and $\boldsymbol{w}$ are the weights of the model. In practice, the data samples are batched, and stochastic gradient descent is used.

Although the weights $\boldsymbol{w}$ were tuned to minimize the loss value, the final model is chosen where the weighted f_1 score on the validation split is the highest, since that's the metric that we report.

4 Experiments

4.1 Datasets and Evaluation Metrics

Dataset	**number of dialogues (utterances)**		
	train	val	test
MELD	1,038 (9,989)	114 (1,109)	280 (2,610)
IEMOCAP	100 (4,778)	20 (980)	31 (1,622)

Dataset	**mean number of utterances per dialogue (std.)**		
	train	val	test
MELD	9.6 (5.8)	9.7 (5.4)	9.3 (5.7)
IEMOCAP	47.78 (16.47)	49.0 (17.44)	52.32 (17.36)

Table 1: The upper half of the table shows the number of dialogues and utterances of the datasets.[2]. The bottom half shows the mean and the standard deviation of the number of utterances per dialogue. As shown, IEMOCAP has about 5 times more utterances per dialogue than MELD.

We test EmoBERTa on the two popular ERC datasets. **MELD** (Poria et al., 2019) is a multimodal (visual, audio, and text) and multi-party (more than two interlocutors in a dialogue) conversational dataset. It was collected from the TV series *Friends*. The seven emotions are `neutral`, `joy`, `surprise`, `anger`, `sadness`, `disgust`, and `fear`. Weighted f_1 score is used to evaluate the performance, as the class distribution is highly imbalanced.

IEMOCAP (Busso et al., 2008) is a multimodal (visual, audio, and text) and dyadic (only two interlocutors in a dialogue) conversational dataset. Ten actors participated in the data collection. Although the original dataset contains 11 different emotions, only six of them are used for evaluation. They are `neutral`, `frustration`, `sadness`, `anger`, `excited`, and `happiness`. As MELD, the class distribution is highly imbalanced and thus a weighted f_1 score will be used for evaluation. Unlike MELD, IEMOCAP does not officially have the names of the speakers. Therefore, we gave each actor a random name. See Appendix A.2 for the details.

Some statistics on train, val, and test splits of both datasets can be found at Table 1.

4.2 Baselines

We compare our model with the models we mentioned in Section 2.

5 Results and Analysis

5.1 Quantitative Analysis

Table 2 shows the performance of our models and the baselines. Our models outperform the other models on both MELD and IEMOCAP.

Model		MELD	IEMOCAP
BERT+MTL (Li et al., 2020b)		61.90	-
BiERU-lc (Li et al., 2020c)		60.84	64.65
DialogueGCN (Ghosal et al., 2019)		58.1	64.18
RGAT (Ishiwatari et al., 2020)		60.91	65.22
CESTa (Wang et al., 2020)		58.36	67.1
VHRED (Hazarika et al., 2021)		-	58.6
SumAggGIN (Sheng et al., 2020)		58.45	66.61
COSMIC (Ghosal et al., 2020)		65.21	65.28
KET (Zhong et al., 2019b)		58.18	59.56
BiF-AGRU (Jiao et al., 2019)		58.1	63.5
Iterative (Lu et al., 2020)		60.72	64.37
HiTrans (Li et al., 2020a)		61.94	64.5
DialogXL (Shen et al., 2020)		62.41	65.94
EmoBERTa	No past and future utterances	63.46	56.09
	Only past utterances	64.55	**68.57**
	Only future utterances	64.23	66.56
	Both past and future utterances	**65.61**	67.42
	→ *without speaker names*	65.07	64.02

Table 2: All of the reported values are weighted f_1 (%) score on the test splits. The best model and the best performance values are in bold. Since the values are stochastic in nature, we report the mean values of five random seeds.

EmoBERTa shows very good results: max weighted f_1 scores (%) of 65.61 (MELD) and 68.57 (IEMOCAP) respectively and above the best reported SOTA, especially considering that no modifications were made to the original RoBERTa model architecture. We also trained a model without the speaker names prepended, which drops the performance: weighted f_1 scores (%) of 65.07 (MELD) and 64.02 (IEMOCAP) respectively, pro-

[2] The IEMOCAP dataset does not officially have train, val, and test splits. Therefore, we follow the splits used by (Zhong et al., 2019b), as these splits are widely used.

viding evidence that encoding the speaker information helps.

Especially on IEMOCAP dataset, although we had to come up with random names for the actors, EmoBERTa was able to learn what's important. As for IEMOCAP, the results were better using only past utterances, rather than using both past and future utterances. We believe that this is due to the fact that IEMOCAP has many more utterances than MELD, and thus we couldn't fit all the past and future utterances in one sequence, meaning that only using past utterances can fit more past utterances than aiming for both. Past utterances were apparently more useful than future utterances to predict the emotion of a current utterance.

There was a bigger performance gain by incorporating past and/or future utterances in IEMOCAP than MELD (12.48 vs. 2.15). We believe that this is due to the fact that the nature of IEMOCAP is more contextual than that of MELD.

5.2 Qualitative Analysis

To get more insight in the value of encoding the speaker, we did a qualitative analysis on 10 correctly and 10 incorrectly classified random samples from each test split.

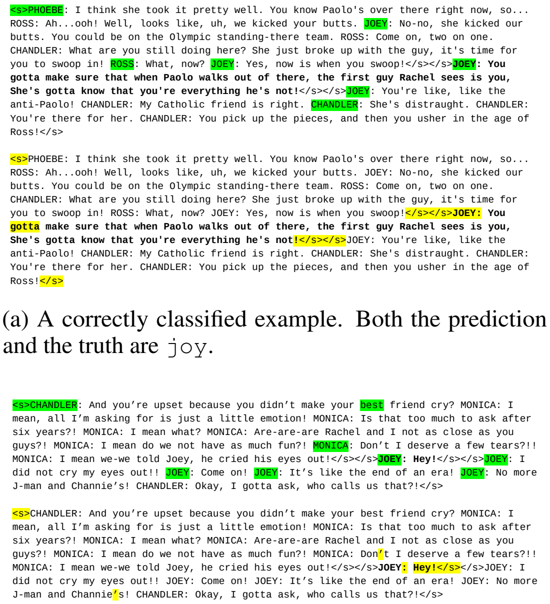
Our manual inspection shows that the model tries to learn the dynamics of the interlocutors in the beginning layer, as the current speaker tokens attend to the interlocutor tokens (We observed this behavior from all the 20 MELD and 20 IEMOCAP random test samples). As for MELD, in all 100% of the correctly classified samples, based on the top 10 attended tokens, the <s> token of the last layer attended to the speaker token of the target (current) utterance, whereas this ratio was only 60% for the incorrectly classified samples. This verifies that the current speaker token increasingly contains important information, as the tokens move on to the higher layers.

Figure 1 shows a visualization of one correctly and one incorrectly classified random samples.

As the <s> token in the last layer focuses on the current speaker and his/her utterance, we believe that this information is what the model finds most useful to make the final prediction.

Note that in the incorrectly classified example, the <s> token in the last layer does not focus on the current speaker but some random punctuation marks throughout the conversation, thus leading to an incorrect prediction.

See Appendix A.3 for the random IEMOCAP test samples.



<s>PHOEBE: I think she took it pretty well. You know Paolo's over there right now, so... ROSS: Ah...ooh! Well, looks like, uh, we kicked your butts. JOEY: No-no, she kicked our butts. You could be on the Olympic standing-there team. ROSS: Come on, two on one. CHANDLER: What are you still doing here? She just broke up with the guy, it's time for you to swoop in! ROSS: What, now? JOEY: Yes, now is when you swoop!</s></s>**JOEY: You gotta make sure that when Paolo walks out of there, the first guy Rachel sees is you, She's gotta know that you're everything he's not!**</s></s>JOEY: You're like, like the anti-Paolo! CHANDLER: My Catholic friend is right. CHANDLER: She's distraught. CHANDLER: You're there for her. CHANDLER: You pick up the pieces, and then you usher in the age of Ross!</s>

<s>PHOEBE: I think she took it pretty well. You know Paolo's over there right now, so... ROSS: Ah...ooh! Well, looks like, uh, we kicked your butts. JOEY: No-no, she kicked our butts. You could be on the Olympic standing-there team. ROSS: Come on, two on one. CHANDLER: What are you still doing here? She just broke up with the guy, it's time for you to swoop in! ROSS: What, now? JOEY: Yes, now is when you swoop!</s></s>**JOEY: You gotta make sure that when Paolo walks out of there, the first guy Rachel sees is you, She's gotta know that you're everything he's not!**</s></s>JOEY: You're like, like the anti-Paolo! CHANDLER: My Catholic friend is right. CHANDLER: She's distraught. CHANDLER: You're there for her. CHANDLER: You pick up the pieces, and then you usher in the age of Ross!</s>

(a) A correctly classified example. Both the prediction and the truth are joy.

<s>CHANDLER: And you're upset because you didn't make your best friend cry? MONICA: I mean, all I'm asking for is just a little emotion! MONICA: Is that too much to ask after six years?! MONICA: I mean what? MONICA: Are-are-are Rachel and I not as close as you guys?! MONICA: I mean do we not have as much fun?! MONICA: Don't I deserve a few tears?!! MONICA: I mean we-we told Joey, he cried his eyes out!</s></s>**JOEY: Hey!**</s></s>JOEY: I did not cry my eyes out!! JOEY: Come on! JOEY: It's like the end of an era! JOEY: No more J-man and Channie's! CHANDLER: Okay, I gotta ask, who calls us that?!</s>

<s>CHANDLER: And you're upset because you didn't make your best friend cry? MONICA: I mean, all I'm asking for is just a little emotion! MONICA: Is that too much to ask after six years?! MONICA: I mean what? MONICA: Are-are-are Rachel and I not as close as you guys?! MONICA: I mean do we not have as much fun?! MONICA: Don't I deserve a few tears?!! MONICA: I mean we-we told Joey, he cried his eyes out!</s></s>**JOEY: Hey!**</s></s>JOEY: I did not cry my eyes out!! JOEY: Come on! JOEY: It's like the end of an era! JOEY: No more J-man and Channie's! CHANDLER: Okay, I gotta ask, who calls us that?!</s>

(b) An incorrectly classified example. The prediction is joy while the truth is anger.

Figure 1: Two examples from the 20 randomly selected test samples are shown. The current speaker utterance, of which the emotion that the model has to predict, is in bold. The green highlighted tokens are the top 10 most attended tokens to the current speaker (i.e., JOEY) in the beginning layer of the model. The yellow highlighted tokens are the top 10 most attended tokens to the [CLS] token (i.e., <s>) in the last layer.[3]. Best viewed when zoomed in.

6 Conclusion

In this paper, we showed that our new model, EmoBERTa, outperforms other models in the ERC task. Since EmoBERTa can directly attend to the input tokens and interlocutor names, we can easily observe the attention coefficients to see which part of the dialogue the model finds most important to make a final classification.

Acknowledgements

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture

[3] Since there are 16 attention heads used per layer in RoBERTa, the visualized weight coefficients are the mean values of them. Since RoBERTa uses a BPE tokenizer (Sennrich et al., 2016), the speaker names (e.g., JOEY) are often separated into more than one token (e.g., JO and EY). Therefore, we highlight the full names, even though only some parts of them are highlighted.

and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

C. Busso, M. Bulut, Chi-Chun Lee, Ebrahim Kazemzadeh, Emily Mower Provost, S. Kim, J. N. Chang, S. Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *CoRR*, abs/1908.11540.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677.

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

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. 2019. Real-time emotion recognition via attention gated hierarchical memory network. *CoRR*, abs/1911.09075.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Anders Krogh and John A. Hertz. 1991. A simple weight decay can improve generalization. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, page 950–957, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020a. 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.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020b. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *ArXiv*, abs/2003.01478.

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

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. Cite arxiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4078–4088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2017. Mixed precision training. *CoRR*, abs/1710.03740.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *CoRR*, abs/2012.08695.

Dongming Sheng, Dong Wang, Ying Shen, Haitao Zheng, and Haozhuang Liu. 2020. Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4153–4163, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019a. Knowledge-enriched transformer for emotion detection in textual conversations. *CoRR*, abs/1909.10681.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019b. 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.

A Appendix

A.1 Training Details

We used GCP (Google Cloud Platform) Compute Engine to carry out our experiments. We used an NVIDIA Tesla V100 machine (disclaimer: we are supported by neither Google nor NVIDIA). This GPU has 16 GB of memory and depending on the length of the input sequence, we were able to fit 4 to 16 samples in one batch, using mixed precision. The pretrained `roberta-large` model has about 355 million parameters. We found that without mixed precision, it's very difficult to train this model, since it's a pretty big model.

We set the value of L^2 regularization rate as 0.01. Training was done for five epochs. No weights were frozen during the training. The learning rate scheduler was set to linearly increase in the first 20% of training and then linearly decrease in the remaining 80%.

Since the optimal peak learning rate highly depends on the batch size and the other hyperparameters, we used Optuna (Akiba et al., 2019) to find its best value. 10% of the training data and the same amount of validation data were used to search for the best value. Optuna ran five trials and looked for the best learning rate, between $1e-6$ and $1e-4$, that minimizes the cross entropy loss on the validation data split.

The hyperparameters not mentioned here are all set to the default values.

One full five-epoch training took about 45 minutes.

A.2 IEMOCAP Speaker Names

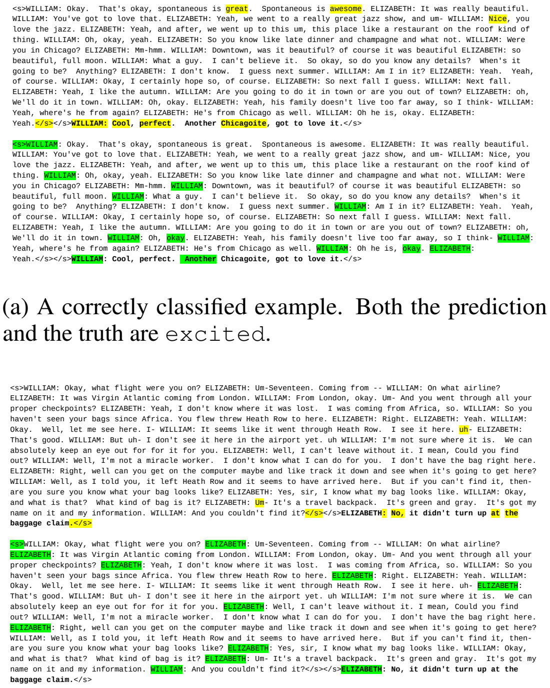
Since the IEMOCAP dataset was created in the US, we used the top five male and female American names over the past 100 years (https://www.ssa.gov/oact/babynames/decades/century.html).

The female names used are `Mary`, `Patricia`, `Jennifer`, `Linda`, and `Elizabeth`. The male names used are `James`, `John`, `Robert`, `Michael`, and `William`.

A.3 Qualitative Analysis on IEMOCAP

Unlike MELD, only 20% of the correctly classified samples showed the behavior of the last layer `<s>` token attending to the current target speaker. This ratio was 10% for the incorrectly classified samples. We believe this is due to the fact that the speaker names in the test split of IEMOCAP (i.e., `WILLIAM` and `ELIZABETH`) were never seen during training.)

Figure 2 gives you a visualization of the qualitative analysis on the IEMOCAP dataset.



<s>WILLIAM: Okay. That's okay, spontaneous is great. Spontaneous is awesome. ELIZABETH: It was really beautiful. WILLIAM: You've got to love that. ELIZABETH: Yeah, we went to a really great jazz show, and um- WILLIAM: Nice, you love the jazz. ELIZABETH: Yeah, and after, we went up to this um, this place like a restaurant on the roof kind of thing. WILLIAM: Oh, okay, yeah. ELIZABETH: So you know like late dinner and champagne and what not. WILLIAM: Were you in Chicago? ELIZABETH: Mm-hmm. WILLIAM: Downtown, was it beautiful? of course it was beautiful ELIZABETH: so beautiful, full moon. WILLIAM: What a guy. I can't believe it. So okay, so do you know any details? When's it going to be? Anything? ELIZABETH: I don't know. I guess next summer. WILLIAM: Am I in it? ELIZABETH: Yeah. Yeah, of course. WILLIAM: Okay, I certainly hope so, of course. ELIZABETH: So next fall I guess. WILLIAM: Next fall. ELIZABETH: Yeah, I like the autumn. WILLIAM: Are you going to do it in town or are you out of town? ELIZABETH: oh, We'll do it in town. WILLIAM: Oh, okay. ELIZABETH: Yeah, his family doesn't live too far away, so I think- WILLIAM: Yeah, where's he from again? ELIZABETH: He's from Chicago as well. WILLIAM: Oh he is, okay. ELIZABETH: Yeah.</s></s>**WILLIAM: Cool, perfect. Another Chicagoite, got to love it.**</s>

<s>WILLIAM: Okay. That's okay, spontaneous is great. Spontaneous is awesome. ELIZABETH: It was really beautiful. WILLIAM: You've got to love that. ELIZABETH: Yeah, we went to a really great jazz show, and um- WILLIAM: Nice, you love the jazz. ELIZABETH: Yeah, and after, we went up to this um, this place like a restaurant on the roof kind of thing. WILLIAM: Oh, okay, yeah. ELIZABETH: So you know like late dinner and champagne and what not. WILLIAM: Were you in Chicago? ELIZABETH: Mm-hmm. WILLIAM: Downtown, was it beautiful? of course it was beautiful ELIZABETH: so beautiful, full moon. WILLIAM: What a guy. I can't believe it. So okay, so do you know any details? When's it going to be? Anything? ELIZABETH: I don't know. I guess next summer. WILLIAM: Am I in it? ELIZABETH: Yeah. Yeah, of course. WILLIAM: Okay, I certainly hope so, of course. ELIZABETH: So next fall I guess. WILLIAM: Next fall. ELIZABETH: Yeah, I like the autumn. WILLIAM: Are you going to do it in town or are you out of town? ELIZABETH: oh, We'll do it in town. WILLIAM: Oh, okay. ELIZABETH: Yeah, his family doesn't live too far away, so I think- WILLIAM: Yeah, where's he from again? ELIZABETH: He's from Chicago as well. WILLIAM: Oh he is, okay. ELIZABETH: Yeah.</s></s>**WILLIAM: Cool, perfect. Another Chicagoite, got to love it.**</s>

(a) A correctly classified example. Both the prediction and the truth are `excited`.

<s>WILLIAM: Okay, what flight were you on? ELIZABETH: Um-Seventeen. Coming from -- WILLIAM: On what airline? ELIZABETH: It was Virgin Atlantic coming from London. WILLIAM: From London, okay. Um- And you went through all your proper checkpoints? ELIZABETH: Yeah, I don't know where it was lost. I was coming from Africa, so. WILLIAM: So you haven't seen your bags since Africa. You flew threw Heath Row to here. ELIZABETH: Right. ELIZABETH: Yeah. WILLIAM: Okay. Well, let me see here. I- WILLIAM: It seems like it went through Heath Row. I see it here. uh- ELIZABETH: That's good. WILLIAM: But uh- I don't see it here in the airport yet. uh WILLIAM: I'm not sure where it is. We can absolutely keep an eye out for for it for you. ELIZABETH: Well, I can't leave without it. I mean, Could you find out? WILLIAM: Well, I'm not a miracle worker. I don't know what I can do for you. I don't have the bag right here. ELIZABETH: Right, well can you get on the computer maybe and like track it down and see when it's going to get here? WILLIAM: Well, as I told you, it left Heath Row and it seems to have arrived here. But if you can't find it, then- are you sure you know what your bag looks like? ELIZABETH: Yes, sir, I know what my bag looks like. WILLIAM: Okay, and what is that? What kind of bag is it? ELIZABETH: Um- It's a travel backpack. It's green and gray. It's got my name on it and my information. WILLIAM: And you couldn't find it?</s></s>**ELIZABETH: No, it didn't turn up at the baggage claim.**</s>

<s>WILLIAM: Okay, what flight were you on? ELIZABETH: Um-Seventeen. Coming from -- WILLIAM: On what airline? ELIZABETH: It was Virgin Atlantic coming from London. WILLIAM: From London, okay. Um- And you went through all your proper checkpoints? ELIZABETH: Yeah, I don't know where it was lost. I was coming from Africa, so. WILLIAM: So you haven't seen your bags since Africa. You flew threw Heath Row to here. ELIZABETH: Right. ELIZABETH: Yeah. WILLIAM: Okay. Well, let me see here. I- WILLIAM: It seems like it went through Heath Row. I see it here. uh- ELIZABETH: That's good. WILLIAM: But uh- I don't see it here in the airport yet. uh WILLIAM: I'm not sure where it is. We can absolutely keep an eye out for for it for you. ELIZABETH: Well, I can't leave without it. I mean, Could you find out? WILLIAM: Well, I'm not a miracle worker. I don't know what I can do for you. I don't have the bag right here. ELIZABETH: Right, well can you get on the computer maybe and like track it down and see when it's going to get here? WILLIAM: Well, as I told you, it left Heath Row and it seems to have arrived here. But if you can't find it, then- are you sure you know what your bag looks like? ELIZABETH: Yes, sir, I know what my bag looks like. WILLIAM: Okay, and what is that? What kind of bag is it? ELIZABETH: Um- It's a travel backpack. It's green and gray. It's got my name on it and my information. WILLIAM: And you couldn't find it?</s></s>**ELIZABETH: No, it didn't turn up at the baggage claim.**</s>

(b) An incorrectly classified example. The prediction is `neutral` while the truth is `frustration`.

Figure 2: Two examples from the 20 randomly selected test samples are shown. The current speaker utterance, of which the emotion that the model has to predict, is in bold. The green highlighted tokens are the top 10 most attended tokens to the current speaker (i.e., `WILLIAM` and `ELIZABETH`, for Figure 2a and 2b, respectively.) in the beginning layer of the model. The yellow highlighted tokens are the top 10 most attended tokens to the `[CLS]` token (i.e. `<s>`) in the last layer. Unlike Figure 1, there is only one `[SEP]` token (i.e., `</s></s>`), since this model only has two segments, past and current. Best viewed when zoomed in

We see a similar behavior as in MELD. Again, in the incorrectly classified example, the `<s>` in the last layer does not attend to the current speaker.