Multilingual Emotion Recognition in Conversation

Junyin Chen Hanshu Ding Zoe Fang Yifan Jiang
{junyinc, hsding99, zoekfang, yfjiang}@uw.edu
Department of Linguistics
University of Washington

Abstract

TBD

1 Introduction

TBD

2 Task Description

2.1 Primary Task

Our primary task is emotion recognition in conversation (ERC) task on the text modality of the Multimodal EmotionLines Dataset (MELD). The dataset is in English and contains dialogues and utterances from TV series scripts . We will predict the emotion for each utterance from dialogues involving multiple speakers.

2.1.1 Primary Dataset

MELD (Poria et al., 2019) [1], also kown as *Multimodel Emotionlines Dataset*, is a multi-party emotional conversational database that is extended from Emotionlines dataset. Emotionlines (Hsu et al., 2018)[2] dataset is an emotion dialogue dataset with emotion labels for each utterance. The utterances are collected from Friends TV scripts and private Facebook messenger dialogues. Each utterance is labeled with one of Ekman's six basic emotions plus the neutral emotion. MELD is an upgraded emotion dataset that contains about 13,000 utterances from 1,433 dialogues from only the TV-series Friends. Each utterance in MELD is annotated with emotion and sentiment labels, and encompasses audio, visual, and textual modalities. MELD splits the data into training, development, and testing set separately (1039 dialogues for training set, 114 dialogues for developing set and 280 dialogues for testing set). We use the testing set and corresponding gold standard annotation for analysis.

According to Poria, the speakers of these dialogues are categorized to 6 main characters and others. The utterances are distributed relatively evenly with respect to speakers (from 12% to 16%). The emotion distribution for each character is also similar.

We also use the EmoryNLP (Zahiri and Choi, 2017) as an auxiliary dataset to facilitate the training. EmoryNLP is another multi-party emotional conversational database curated by Zahiri and Choi. The utterances are also collected from the Friends TV show and are annotated with emotion and speaker labels by crowdsourced workers based on Willcox's feeling wheel. EmoryNLP splits the dataset into training, development, and testing sets separately. We use the testing set and corresponding gold standard annotation for analysis.

According to Zahiri and Choi, the distribution of all emotions in the corpus is not even. The two most dominant emotions, *neutral* and *joyful*, together yield over 50% of the dataset.

2.2 Adaptation Task

Our adaptation task is to adapt our model to Chinese dialogues in the Multi-party Dialogue Dataset (MPDD). Other dimensions for this task remain the same as the primary task.

2.2.1 Dataset

Multi-party Dialogue Dataset (MPDD) (Chen et al., 2020) [3] is a Chinese emotional conversational dataset. The dataset contains a total of 25, 548 utterances from 4, 142 dialogues, which are collected from five TV series scripts from www.juban108.com. Each uttrance is annmotated with three types of labels: emotion, relation, and target listener. In particular, the emotion labels are consistent with those in the Emotionlines dataset.

[1] https://affective-meld.github.io/

[2] http://doraemon.iis.sinica.edu.tw/emotionlines/index.html

[3] http://nlg.csie.ntu.edu.tw/nlpresource/MPDD

3 System Overview

3.1 Design

Instead of appending the correspondent speaker tag with the utterance for training, we integrate the ideology of having a multi-task deep neural network that shares the lower layer across multiple single-sentence and pairwise text classification tasks.

With the benefit of multi-task deep neural network, we enable our model to gain context awareness with multiple context related tags, such as speaker, past and future utterance.

3.2 Methodology

We want to determine the emotion of a specific utterance in a dialogue group M, which contains multiple speakers in speaker group S. The dialogue group M can be expressed as a list of vectors: $dialogue = [x_1, x_2, ..., x_{n-1}, x_n]$, where each utterance contains multiple words. The correspondent speaker list can be expressed as another list of vectors: $speaker = [s_1, s_1, ..., s_2, s_3]$, where s_n is in the speaker group S. Since this is a supervised setup, we will utilize the manually labeled emotion label e_n corresponding to each utterance x_n in the dialogue group M.

The simplest solution is to have a function f that takes each utterance x_t as input and output the correct label e_t. For our setup, we decide on a multi-tasking function f to output correspondent speaker s_t, as we want to take the speaker into account. Furthermore, we add past utterance $[x_1, x_2, ..., x_{t-1}]$ and even future utterance $[x_{t+1}, x_{t+2}, ..., x_n]$, as we anticipate adding more context around the utterance for analysis will further improve the result.

3.3 Algorithm

For each task in the multi-task deep neural network, we use the pretrained "Roberta-base" model Liu et al. (2019). We choose BERT like algorithms as they both have simple structure, and support more than one segment for tokenization. We choose to use RoBERTa as the main algorithm for training, facilitating result comparison with results listed in Kim and Vossen (2021). Even though pre-trained BERT and RoBERTa models do not expect more than two segments as inputs, both Kim et al. and us show that having more than two segments improves evaluation results.

We will have either two or three segments, if we want to include both past and future utterances.

4 Approach

4.1 Problem Statement

MELD provides information on the speaker and turn ID (in the dialogue) of each utterance. We want to take these two factors into account and build a model that learns emotion prediction based on the context and/or speaker information.

4.2 Multi-Task Deep Neural Network

We chose to build a multi-task deep neural network. The shared layer is a RoBERTa base encoder. For the different task heads, we set the main task to be emotion recognition on the MELD data-set, which is also the data-set we run our final evaluation on. The auxiliary task trains parallel to the main task, the point of the auxiliary task is for the model to learn relevant information in addition to the main task. In our model, we have two options for auxiliary task. One of the options is speaker classification, since past literature on emotion recognition in conversation suggest that the models might perform better when they learn speaker-specific features, and we think there might be correlation between each speaker and a certain emotion distribution. The other option for auxiliary task is that we can use data augmentation and run the same emotion classification task, but on additional data-set (EmoryNLP). We can also have more than one auxiliary task, so we run both speaker classification and data augmentation.

4.3 Input Preprocessing

Other than the model architecture, our approach also involves specific techniques in input preprocessing. The default input is an utterance in a dialogue, given the turn ID, which is the sequential number in the dialogue, and the speaker information. Since we want to experiment on the impact of speaker information on model performance, we can add a speaker classification auxiliary task, or we can do this in input preprocessing, where we provide the speaker information by concatenating them to the utterance, so that the speaker can potentially be a hint for the model to learn. Another way to preprocess input is to provide context of the utterance. We can provide a certain amount of past or future utterances of the utterance that we want to predict. The caveat of training a MT-DNN on top of RoBERTa is that BERT-like models are previously trained with inputs no longer than two sentences, where each two sentence has a BOS

<s> token and an EOS </s> token. For that reason, when we concatenate more than one past utterance to the to-predict utterance, we strip the EOS token of the past utterances and keep only one that exists between the context and the to-predict utterance. Algorithm 1 shows our method of adding context and speaker to an utterance.

Algorithm 1 Add All Past Utterance(s)

Given an $Uttr$ and its idx in dialogue
if $idx \neq 0$ **then**
  $i \leftarrow 1$
  while $idx - i \geq 0$ **do**
    $Uttr_{past} \leftarrow Data_{idx-i}$
    $Speaker_{past} \leftarrow Uttr_{Past_{speaker}}$
    $str \leftarrow Speaker_{past} + Uttr_{Past}$
    $Context \leftarrow Context + str$
    $i \leftarrow i + 1$
$Uttr \leftarrow Context + EOS + Uttr$

4.4 Evaluation

Both tasks will be evaluated using weighted F1 to account for the imbalance of the dataset. We will also evaluate the error rate between predicted the true emotion labels that are labelled manually by the dataset curators.

5 Results

past utterance	future utterance	weighted F1
0	0	60.17
0	6	61.96
6	0	62.46

Table 1: Weighted F1 when number of past or future utterances are added as context. Auxiliary task = Speaker Classification. Speaker in input = True.

speaker task	speaker in input	weighted F1
Yes	No	63.75
Yes	Yes	62.92
No	Yes	62.60
No	No	64.74

Table 2: Weighted F1 when auxiliary task is speaker classification and/or speaker information is concatenated to input utterance. Number of Past Utterances = 10

Using our evaluation metric, the model currently yields a highest 64.74 percent weighted F1 score. Table 1 shows the results before we add the EmoryNLP dataset, and we have speaker classification task as the auxiliary task. Table 1 shows that adding only past utterances yields the best result. Table 2 shows the results when we set the one of the auxiliary tasks to emotion prediction, while keeping the second auxiliary task of speaker classification optional. We also experiment with the option of having speaker in input.

emotion	true count	total count	accuracy
neutral	1026	1256	0.8169
joy	247	402	0.6144
anger	156	345	0.4522
surprise	173	281	0.6157
sadness	66	208	0.3173
disgust	5	68	0.0735
fear	1	50	0.02

Table 3: Prediction accuracy for each emotion in model with speaker classification task

emotion	true count	total count	accuracy
neutral	985	1256	0.7842
joy	263	402	0.6542
anger	173	345	0.5014
surprise	180	281	0.6429
sadness	76	208	0.3654
disgust	13	68	0.1912
fear	11	50	0.22

Table 4: Prediction accuracy for each emotion in best-performing model

The prediction output and accuracy is shown as above in Table 3 and 4. A major trend we see from the tables is that the prediction for "neutral" has the highest accuracy compared to other emotions. Before data augmentation, the accuracy for the prediction of specific emotion decreases as the quantity of learning instances for the corresponding emotion decreases. After data augmentation, distributional bias in the data-set is not as patent. Specifically, we see a great improvement in "fear" prediction.

6 Discussion

6.1 Pre-trained model decision

Using the RoBERTA-base pre-trained model, we achieve a similar result illustrated in Kim et al. However, using the RoBERTa-large or RoBERTa-full pre-trained model used by Kim et al., weighted

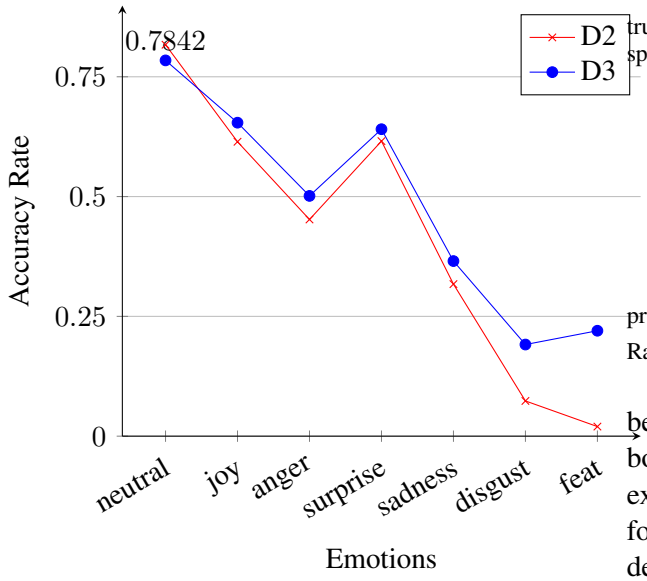

Figure 1: D2 Accuracy vs D3 Accuracy

F1 score reduced by as much as 30 percent, regardless of the number of utterances or auxiliary tasks included in the training dataset. We choose to stick with the RoBERTa-base model for Delivery 3.

6.2 Number of past and future utterance.

Adding future utterance alone does not greatly improve the result. The result of adding future utterance alone is similar of having no future or past utterance. Adding both past and future utterance, preforms almost the same as adding only past utterance. Due to GPU memory size, we choose to only add past utterance in Delivery 3 and increase number of include past utterance to 10.

6.3 Qualitative Analysis

We did a qualitative analysis with on 10 correctly and 10 incorrectly classified random sample from the test split for Delivery 2. For all the incorrectly classified random sample, the speaker tags are all incorrectly classified. For the 10 correctly classified random sample, 40 percent of the speaker tags are correctly classified. This shows that having speaker tag does help improve the result, but not as substantial as we hopped.

> <s>Rachel:Oh, that sounds great.Others:How does that sound?Others:Well, I've got a project for you that's a lot more related to fashion.Others:Well, don't think I haven't noticed your potential.Rachel:Oh, you got me.Others:Eh.</s></s>Come on over here, sweetheart.</s>

A correctly classified example. Both the prediction and truth are 'neutral'. Both the predicted speaker and the true speaker are Others.

> <s>Monica:Is that too much to ask after six year?!Monica:I mean, all I'm asking for is just a little emotio! Chandler:And you're upset because you didn't make your best friend cry?</s></s>I mean what?</s>

An incorrectly classified example. The prediction is 'surprise' while the truth is 'anger'. The predicted speaker is Rachel while the true speaker is Monica.

We initially anticipate that the speaker tag will be utilized for sentiment classification. However, both the incorrectly and correctly classified example, the <s> token in the last layer mainly focus on punctuation marks. This phenomenon far deviates from Kim et al.'s result, that <s> token mainly focus on the speaker in correctly classified examples. This proves that separating speaker and utterance for speaker detection does not benefit the model to utilize speaker context for sentiment analysis.

For Delivery 3, we did another qualitative analysis with ten correctly and ten incorrectly classified random samples from the MELD test split.

> <s>Joey:You know, I think I was sixteen. Monica:Please, just a little bit off the back. Phoebe:Iḿ still on "no." </s></s>Uh, morning. Do you guys think you could close your eyes for just a sec?</s>

A correctly classified example when training with speaker context. Both the prediction and truth are 'neutral'. The predicted speaker is Monica, and the true speaker is Rachel.

> <s>Rachel:It's not a purse! It's a shoulder bag. Joey:It looks like a women's purse. Rachel:No Joey, look. Trust me, all the men are wearing them in the spring catalog. Look. See look, Joey:See look, </s></s>Exactly! Unisex!</s>

An incorrectly classified example when training with speaker context. The prediction is 'joy' while the truth is 'neutral'. The predicted and the true speaker is Rachel.

After using EmoryNLP for auxiliary emotion and speaker detection tasks, we notice an increase in accuracy in the emotion detection task. However, the speaker contexts are mainly utilized by incorrect outputs. Only 5 of the correct samples utilize the speaker tag, whereas all incorrect samples utilize the speaker tag. Both correct and incorrect samples have sixty percent accuracy in predicting the speaker. Based on the new finding, we revised our hypothesis, that adding speaker contexts does not improve the emotion detection accuracy. We then train a new model without the speaker context.

<s>You know, I think I was sixteen. Please, just a little bit off the back. Phoebe:Iḿ still on "no." </s></s>Uh, morning. Do you guys think you could close your eyes for just a sec?</s>

A correctly classified example when training without speaker context. Both the prediction and truth are 'neutral'.

<s>It's not a purse! It's a shoulder bag. It looks like a women's purse. No Joey, look. Trust me, all the men are wearing them in the spring catalog. Look. See look, See look, </s></s>Exactly! Unisex!</s>

An incorrectly classified example when training without speaker context. The prediction is 'joy' while the truth is 'neutral'.

We re-examine the selected random samples with the new models, and the result is the same. We then randomly selected 10 correct samples using the new model and process them with the old model. The old model failed to correctly predict three of the utterances.

Our new model, with an weighted F1 of 64.7%, suggesting speaker context is unnecessary when adding additional dataset.

7 Conclusion

TBD

References

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 610–614, Marseille, France. European Language Resources Association.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. *arXiv e-prints*, page arXiv:2108.12009.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. 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.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *CoRR*, abs/1708.04299.