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

We use a multi-task deep learning model to perform emotion recognition in conversation (ERC) task on English and Chinese utterances from TV series scripts. Our experiment shows that adding speaker tags do not improve performance, adding past utterances might add obfuscation, and simple ambiguous utterance is a challenge. Our results show that multilingual training yields better accuracy than monolingual training. We suggest future works to include additional audio data, since the results show that text data limits the interpretability of dialogues.

1 Introduction

Identification of emotions is one of the core tasks in NLP. As a variant of this task, emotion recognition in conversation, henceforward ERC, is a more difficult task, given that dialogues have a different structure than prose, which is what the majority of the language models are based on. Inspired by past works on ERC, we build a model that predicts emotions on an input utterance. More specifically, we build a multi-task learning model to tackle the complex nature of dialogues, for example multi-party participation and contextual information. We want that the multi-task model can help our model learn dialogue-specific features in hopes that this will improve the accuracy on emotion recognition. To further our project, we adapt our model to predict non-English utterance as well.

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.

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

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

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.

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

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

emotion classification task, but on additional dataset (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 Adaption Task

As described in the previous section, our adaptation task is performed on MPDD, which is a dialogue corpus in Mandarin. Our approach to this task is similar to the primary task. We build a Multi-Task Deep Neural Network that sets MPDD as the evaluation dataset. We train a multi-task deep neural network for the prediction. As a continuation of the primary task, we only use data augmentation for the auxiliary tasks. We use the same dataset from the primary task for data augmentation, which is MELD and EmoryNLP. The model uses XLM-RoBERTa-lonformer-base as the encoder, since input of MPDD has an average length that is longer than the limit of a roberta-base encoder.

4.5 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

5.1 Primary results - Monolingual Model

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 yields a highest 64.74 percent weighted F1 score on MELD. 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.

The prediction output and accuracy of our primary task is shown as below 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.

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: MELD 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: MELD Prediction accuracy for each emotion in best-performing model

5.2 Adaptation Task - Multilingual Model

evaluation	training	weighted F1
MELD	zh+en	62.88
MELD	en	61.18
MPDD	zh+en	59.73
MPDD	zh	57.93

Table 5: Weighted F1 for Adaption Task. Shared encoder=xlm-roberta

Table 5 shows the weighted F1 of both monolingual and multilingual models. The results show that multilingual training datasets yield better results than monolingual training datasets.

6 Discussion

6.1 Pre-trained model decision

We choose to use the xlm-roberta-longformer-base model, instead of the RoBERTa model used by Kim et al, due to the utterance length in the MPDD dataset. One utterance in MPDD dataset usually contains several long sentences, whereas the utterance in MELD or EmoryNLP usually contains only one sentence. If we choose to segment sentences in the MPDD utterances, we need to perform sentiment analysis on those newly segmented sentences, which is time-consuming. Instead, we use the xlm-roberta-longformer-base model since it is trained in multiple languages, including Chinese, and can process longer sentences as input.

6.2 Qualitative Analysis

6.2.1 MELD

We did a qualitative analysis with ten correctly and ten incorrectly classified random samples from the evaluation dataset of MELD.

> <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.

We initially anticipate that the speaker tag will be utilized for sentiment classification. However, 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. Our manual inspections show that only 5 of the correct samples focus on the speaker tags in the last layer, whereas all incorrect samples focus on the speaker tags in the last layer. Both correct and incorrect samples have sixty percent accuracy in predicting the speaker. This phenomenon far deviates from Kim et al.'s result, that <s> token mainly focus on the speaker in correctly classified examples. Based on the new finding, we revised

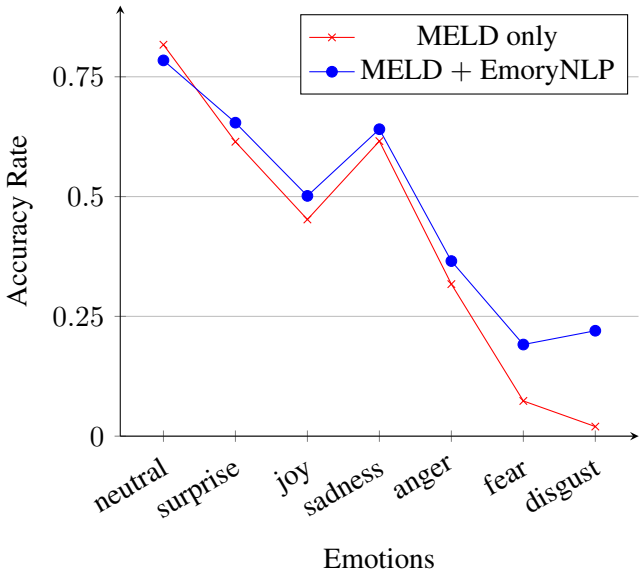

Figure 1: Accuracy Comparison before and after Data Augmentation

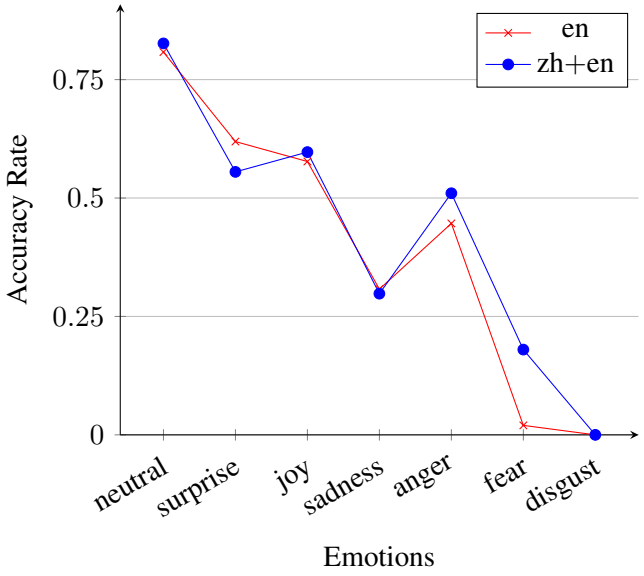

Figure 2: Accuracy Comparison Between English Only Training Set and Multi-lingual Training Set (eval on MELD)

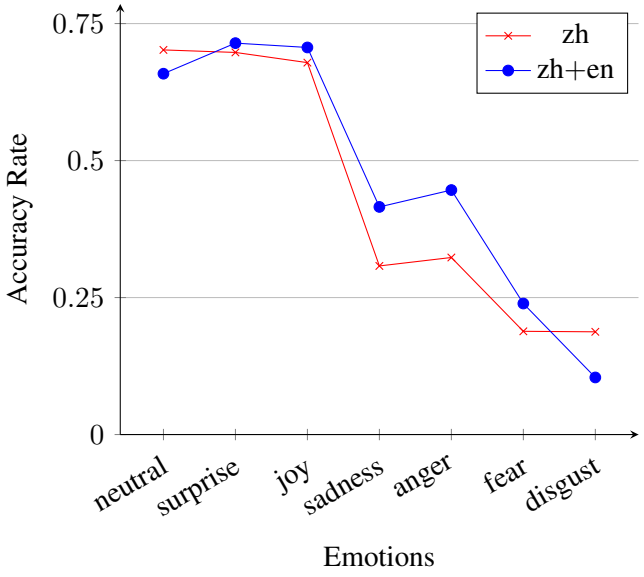

Figure 3: Accuracy Comparison Between English Only Training Set and Multi-lingual Training Set (eval on MPDD)

our hypothesis, that adding speaker contexts does not improve the emotion detection accuracy. We then train a new model without the speaker context.

> <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 without speaker context. Both the prediction and truth are 'joy'.

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

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

Our manual inspections of the new model suggest speaker context is unnecessary when adding a dataset. Nine correct samples and five incorrect samples focus on the previous utterances in the last layer. The incorrect samples usually contain words that are usually associated with the predicted semantic labeling, but when combined with important sentiment keywords, the result becomes opposite. For example, as the incorrect sample listed above, *cry*, *emotion*, and *too* are used in the previous utterances, which can be associated with joy, but with *upset*, the true sentiment becomes sadness.

6.2.2 MPDD

We also did a qualitative analysis with ten correctly and ten incorrectly classified random samples from the evaluation dataset split of the MPDD dataset.

> <s> 應該願意作,比你還願意作。那可不一定,聽說要抽血,可能就不願意了。你跟韓東說了嗎? 說了。光說不行,你得催點,要不得拖到啥時候? 別光說我的事,你也該給你的親爸盡盡孝心了吧? </s></s> 我的親爸?</s>

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

> <s> 八萬是多少? 八萬就是八萬。八萬就八萬吧。咱們就去取吧,在哪兒取? 在鎮政府。那好吧,咱們這就去。要帶上您的身份證和戶口本。</s></s> 行。</s>

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

Seven of the correct sample, and three of the incorrect sample focus on the past utterance in the last layer. Noticeably, the model struggles with short, simple, but sometimes ambiguous utterances. For example, as the incorrect sample listed above, the utterance for judgement is just one Chinese character, 行, which means Okay in English. 行 does carry multiple meanings, but means acknowledging the previous utterance in this case, which the model incorrectly predicted the sentence as joy, while the truth in neutral.

7 Conclusion

In this paper, we showed how to utilize multi-task deep learning model to perform emotion recognition in conversation (ERC) task. Our also showed that multilingual training datasets yield better results than monolingual training dataset. Our manual inspections on model's last layer suggest adding additional textual contexts, such as speaker tags, does not improve the performance. We suggest future works to include additional non-textual contexts, such as audio or video data, since the results show that text data limits the interpretability of dialogues.

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.