applied sciences

MDPI

Article

Hierarchical Transformer Network for Utterance-Level Emotion Recognition

Qingbiao Li, Chunhua Wu *, Zhe Wang and Kangfeng Zheng

School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China; liqingbiao@bupt.edu.cn (Q.L.); wangxiaozhe@bupt.edu.cn (Z.W.); zkf@vip.163.com (K.Z.)

* Correspondence: wuchunhua@bupt.edu.cn; Tel.: +86-186-0010-5255

Received: 10 June 2020; Accepted: 24 June 2020; Published: 28 June 2020

Abstract: While there have been significant advances in detecting emotions in text, in the field of utterance-level emotion recognition (ULER), there are still many problems to be solved. In this paper, we address some challenges in ULER in dialog systems. (1) The same utterance can deliver different emotions when it is in different contexts. (2) Long-range contextual information is hard to effectively capture. (3) Unlike the traditional text classification problem, for most datasets of this task, they contain inadequate conversations or speech. (4) To better model the emotional interaction between speakers, speaker information is necessary. To address the problems of (1) and (2), we propose a hierarchical transformer framework (apart from the description of other studies, the "transformer" in this paper usually refers to the encoder part of the transformer) with a lower-level transformer to model the word-level input and an upper-level transformer to capture the context of utterance-level embeddings. For problem (3), we use bidirectional encoder representations from transformers (BERT), a pretrained language model, as the lower-level transformer, which is equivalent to introducing external data into the model and solves the problem of data shortage to some extent. For problem (4), we add speaker embeddings to the model for the first time, which enables our model to capture the interaction between speakers. Experiments on three dialog emotion datasets, Friends, EmotionPush, and EmoryNLP, demonstrate that our proposed hierarchical transformer network models obtain competitive results compared with the state-of-the-art methods in terms of the macro-averaged F1-score (macro-F1).

Keywords: emotion recognition; text classification; dialog; transformer; pretrained model

1. Introduction

Sentiment analysis, considered one of the most important methods for analyzing real-world communication, is a kind of classification task for extracting emotion from language. It can help us progress in many fields. For example, in the field of data mining, we can research the financial volatility with sentiment analysis, and it can also drive the development of human–computer interactions. In this paper, we consider one of the tasks in this research direction, utterance-level emotion recognition (ULER) [1]. In ULER, an utterance [2] is a unit of a dialog bounded by breathes or pauses, and our goal is to tag each utterance with the indicated emotion (e.g., happy, sad, or angry). Traditional sentiment analysis methods are confined to analyzing only a single sentence or document, regardless of its surrounding information. However, in the field of ULER, contextual information is indispensable in emotional discrimination. When we analyze the emotion of an utterance, the contextual information represents the information from the front and back utterances. For example, in Figure 1, the utterance "Yes, I agree with this point." can deliver different emotions in different contexts. To identify a speaker's emotion precisely, Hazarika et al. [3] proposed contextual representations for prediction with a recurrent neural network (RNN), where each utterance is represented by a feature vector extracted by convolutional neural networks (CNN) at an earlier stage. Similarly, Jiao et al. [4] proposed a hierarchical gated recurrent unit (HiGRU) framework with a lower-level GRU to model the word-level inputs and an upper-level GRU to capture the contexts of utterance-level embeddings. Theoretically, RNNs such as long short-term memory (LSTM) and gated recurrent units (GRUs) should propagate long-term contextual information. However, in practice, this is not always the case [5]. In cases where the input sequence is long, RNNs may experience an exploding gradient or vanishing gradient. Unlike traditional text classification problems, in the field of ULER, there are a limited number of datasets, and most datasets contain inadequate conversations. This issue limits the possibility of obtaining larger models for this task. To solve this issue, Zheng et al. [6] proposed a knowledge enriched transformer (KET) to effectively incorporate contextual information and external knowledge bases, but this model structure is complex, and the running speed is not high. Jiao et al. [7] proposed pretraining a context-dependent encoder (CoDE) for ULER by learning from unlabeled conversation data to address the aforementioned challenge, but the model did not perform better in the word-level embedding phase. Huang and Lee [8] proposed two different bidirectional encoder representation from transformers (BERT) [9] models to deal with two different datasets and performed pretraining with different data sets. They got excellent results in the EmotionX Challenge 2019. However, different pre training methods are used in the two datasets, so the universality of the approach may be weak.

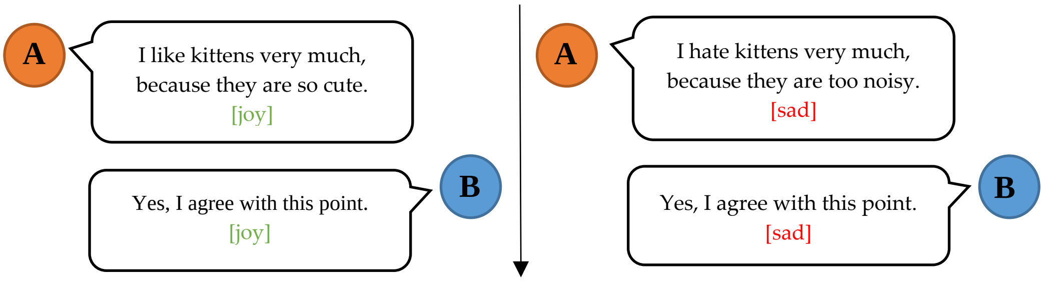
A
I like kittens very much, because they are so cute. [joy]
B
Yes, I agree with this point. [joy]
A
I hate kittens very much, because they are too noisy. [sad]
B
Yes, I agree with this point. [sad]

Figure 1. The utterance "Yes, I agree with this point." can deliver different emotions in different contexts.

In this task, we propose a hierarchical transformer framework to solve the above issues. First, we use a transformer [10] to model the word-level input and capture the contexts of utterance-level embeddings, which has been shown to be a powerful representation learning model in many NLP tasks and can exploit contextual information more efficiently than RNNs and CNNs. Second, for the data scarcity issue, we use a pretrained language model BERT [9] as the lower-level transformer, which is equivalent to introducing external data into the model and helps our model obtain better utterance embedding. Third, the same utterance can deliver different emotions in the same context. For example, in Figure 2, the utterance "Yes, I agree. I think so, too." can deliver different emotions, joy, and sadness. A study also showed that speakers tend to mirror their counterparts to build rapport during the course of a dialogue [11]. Therefore, computational modeling of context should consider emotional dynamics of the interlocutors in a conversation. However, previous studies have not addressed this situation because those models did not capture the interaction between the speakers and did not consider the emotional dynamics of the speakers in a dialog. To solve the problem, we introduce speaker embedding into our model. To the best of our knowledge, this is the first model for ULER with speaker embedding. After obtaining the contextual utterance embedding vectors with a hierarchical transformer framework, we feed them into the fully connected layers for classification. We employ dropout on the fully connected layers to prevent overfitting. Finally, we obtain an utterance category with a softmax layer.

We summarize our contributions as follows:

- We propose a hierarchical transformer framework to better learn both the individual utterance embeddings and the contextual information of utterances.

- We use a pretrained language model, BERT, to obtain better dialog embedding, which is equivalent to introducing external data into the model and solving the problem of data shortage to some extent.
- For the first time, we use speaker embedding in the model for the ULER task, which allows our model to capture the interaction between speakers and better understand emotional dynamics in dialog systems.
- Our model outperforms state-of-the-art models on three benchmark datasets, Friends, EmotionPush, and EmoryNLP.

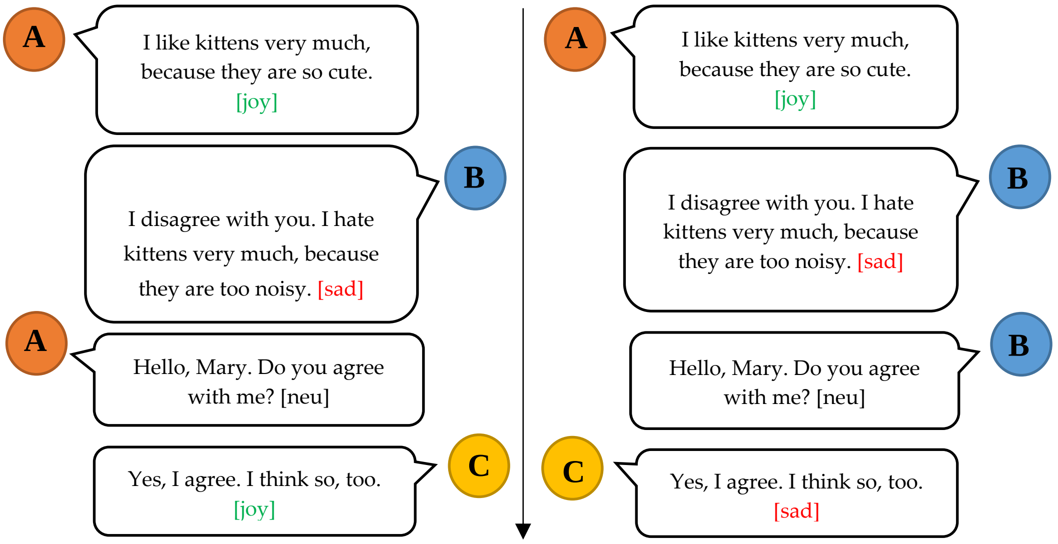
A
I like kittens very much, because they are so cute. [joy]
B
I disagree with you. I hate kittens very much, because they are too noisy. [sad]
A
Hello, Mary. Do you agree with me? [neu]
C
Yes, I agree. I think so, too. [joy]
A
I like kittens very much, because they are so cute. [joy]
B
I disagree with you. I hate kittens very much, because they are too noisy. [sad]
B
Hello, Mary. Do you agree with me? [neu]
C
Yes, I agree. I think so, too. [sad]

Figure 2. The utterance "Yes, I agree. I think so, too." delivers different emotions, joy and sadness, when the previous sentence is from Person A and Person B, respectively.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 describes model architecture. Section 4 outlines the experimental setup. Section 5 discusses the empirical results and analysis. Finally, Section 6 presents the conclusion and future work.

2. Related Work

Text-based emotion recognition is a long-standing research topic, and there have been many excellent studies. In traditional methods, a common method of expressing text is the bag-of-words method. However, the bag-of-words method loses the order of the words. The n-gram model is a very popular statistical language model and usually performs well [12]. However, the n-gram model has a large defect in that it is affected by data sparsity [13]. Recently, neural network methods have become increasingly popular. There is a trend moving from traditional methods to deep learning methods to obtain better text representations. Some prominent models include recursive autoencoders (RAEs) [14], convolutional neural networks (CNNs) [15], and recurrent neural networks (RNNs) [16]. However, these models do not perform well in the field of ULER because they treat texts independently and thus cannot capture the interdependence of utterances in dialogs, or contextual information.

Some studies have tried to solve the above problems. The RNN architecture is a standard method for capturing the sequential relationship of data. Poria et al. [17] captured the contextual information with a bidirectional long short-term memory (BiLSTM) network and obtained great performance. Similarly, Jiao et al. [4] applied bidirectional GRU to model contextual information. In addition, they placed a self-attention layer in the hidden states of GRU and fused the attention outputs with the individual utterance embeddings to learn the contextual utterance embeddings. Luo et al. [18] applied self-attention to model the context of contextual features extracted by BiLSTM. Sayyed et al. [19] proposed sequence-based convolutional neural networks (SCNN) that utilize emotion sequences from previous utterances to detect the emotion of the current utterance. However, when the input sequence is long, RNNs may experience an exploding gradient or vanishing gradient, and we cannot train RNNs in parallel. For those issues, we use a transformer instead of RNNs for feature extraction. The transformer learns the dependencies between words based entirely on self-attention without any recurrent or convolutional layers. Due to its rich representation and fast computation, it has been applied to many NLP tasks, e.g., response matching in dialog systems [20] and language modeling [21]. The success of the transformer has resulted in a large body of follow-up work. Therefore, some transformer variations have also been proposed, such as GPT [22], BERT [9], the universal transformer [23], and CN3 [24].

To solve the problem of data shortage, Jiao et al. [7] proposed pretraining a context-dependent encoder (CoDE) with unlabeled conversation data, but the model did not perform better in the word-level embedding phase. Zhong et al. [6] proposed a knowledge enriched transformer (KET) with external knowledge bases, but this model structure is complex. Unsupervised pretraining is a special case of semisupervised learning where the goal is to find a good initialization point. Pretrained language models, such as ELMo [25], OpenAI GPT [22], and BERT [9], have achieved great success in a variety of NLP tasks, such as sentiment analysis and textual classification. They can generate deep contextualized embeddings since they are pretrained on a massive unlabeled corpus (i.e., English Wikipedia). Some proposed models [26] with pretrained language models have obtained outstanding results on the sentiment analysis task of individual sentences. Nils and Gurevych. Ref. [27] proposed Siamese BERT-networks (SBERT) to obtain sentence embeddings and proved that their model outperforms other state-of-the-art sentence embedding methods. We use a pretrained language model, BERT [9], as the lower-level transformer. In addition, we introduce speaker embedding into our model for the first time, which allows our model to capture the interaction between the speakers.

3. Approach

In this section, we present the task definition and our proposed hierarchical transformer (HiTransformer) network. In addition, we propose a variation in HiTransformer by adding speaker embedding, named HiTransformer-s. The overall architecture of our models is illustrated in Figure 3.

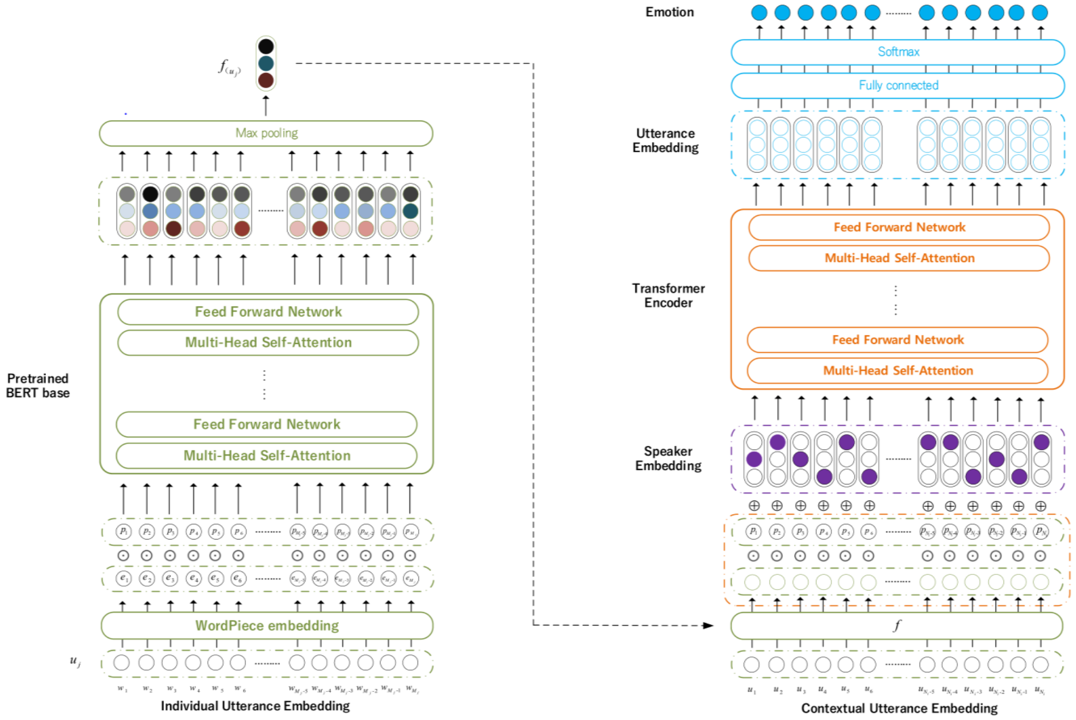

Figure 3. The architecture of our proposed HiTransformer-s. By removing the "Speaker Embedding" layer, we attain the HiTransformer.

3.1. Task Definition

Let there be a set of speakers, $S = \{s_i\}_{i=1}^{M}$, where M is the number of speakers, and a set of emotions, $C = \{c_i\}_{i=1}^{N}$, where N is the number of emotions, such as anger, joy, sadness, and neutral. Assume we are given a set of dialogs, $D = \{D_i\}_{i=1}^{L}$, where L is the number of dialogs. In each dialog, $D_i = \{(u_j, s_j, c_j)\}_{j=1}^{N_i}$ is a sequence of utterances, where the utterance u_j is spoken by $s_j \in S$ with an emotion $c_j \in C$. Our goal is to train a model to find the most likely emotion from C for each new utterance.

3.2. HiTransformer: Hierarchical Transformer

Our HiTransformer consists of two-level transformers: the lower-level transformer models the word-level input and obtains the individual utterance embedding. The upper-level transformer captures the contextual information and obtains contextual utterance embeddings.

3.2.1. Individual Utterance Embedding

For the input utterance $u_j = \{w_k\}_{k=1}^{M_j}$, where u_j is the $j-th.$ utterance in D_i and M_j is the number of words in the utterance u_j. First, the utterance u_j is in lower-case and is tokenized according to a byte pair encoding (BPE) algorithm. If there are tokens exceeding the preset maximum length of input tokens, those tokens are excluded from the list, which may cause information loss and affects the accuracy of the results. Then, we embed those tokens through WordPiece embeddings [28] and obtain the token embeddings $e = \{e_k\}_{k=1}^{M_j}$. Finally, the input embeddings $E = \{E_k\}_{k=1}^{M_j}$ are the summation of the token embeddings e and the positional embeddings $p = \{p_k\}_{k=1}^{M_j}$, which are obtained using the approach used by BERT [9]:

$$E_k = e_k \odot p_k \ (k\epsilon[1, M_j]) \tag{1}$$

where $\odot$ denotes element-wise addition.

We feed the input embeddings E into the lower-level transformer to learn the individual utterance embedding. We adopt the transformer-based pretrained language model BERT (illustrated in Figure 4) as the lower-level transformer, which is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning both the left and right contexts in all layers. The detailed structure is shown in Figure 4. The language model converts input embeddings E into contextual word embedding $T = \{T_k\}_{k=1}^{M_j}$.

$$T = BERT(E) \tag{2}$$

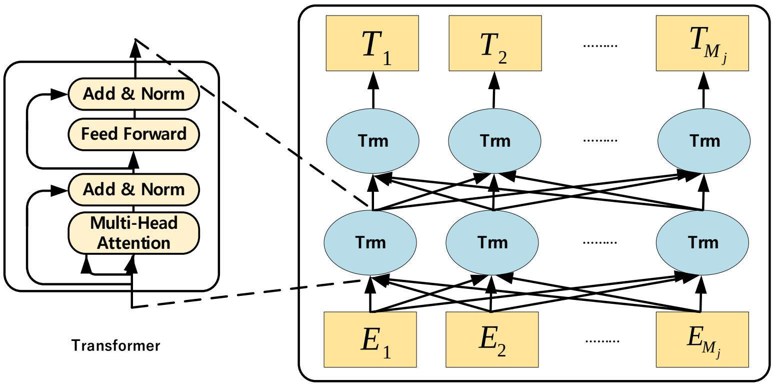

Figure 4. Structure of BERT.

The individual utterance embedding is then obtained by max-pooling of the contextual word embeddings within an utterance, which can assist in retaining important information in each dimension:

$$f(u_j) = maxpool(T) \tag{3}$$

3.2.2. Contextual Utterance Embedding

For the $i-th$ dialog in D, $D_i = \{(u_j,\ s_j,\ c_j)\}_{j=1}^{N_i}$, the individual utterance embedding is $\{f(u_j)\}_{j=1}^{N_i}$. We concatenate the individual embeddings with the position embeddings to obtain $U = \{f(u_j)\ \odot\ p_j\}_{j=1}^{N_i}$, where p_j is the embedding of position j. Then, we feed U into the upper-level transformer to capture the sequential and contextual relationship of utterances in a dialog and obtain the contextual utterance embedding $t = \{t_j\}_{j=1}^{N_i}$.

$$t = Transformer(U) \tag{4}$$

Then, we feed the contextual utterance embedding vector into the classifier, which consists of two linear layers with activation function and dropout. Finally, we obtain the predicted vector over all emotions with a softmax function.

$$\hat{y}_j = softmax(W_2 \cdot Relu(W_1 \cdot t_j + b_1) + b_2) \tag{5}$$

where W_1 and b_1 are the respective weight matrix and bias vector of the first layer, and W_2 and b_2 are the respective weight matrix and bias vector of the second layer.

3.3. *HiTransformer-s: Hierarchical Transformer with Speaker Embeddings*

The HiTransformer has the main issue that it cannot capture the interaction of speakers in a dialog. For example, in Figure 2, the utterance "Yes, I agree. I think so, too." delivers different emotions, sadness and joy. However, the HiTransformer cannot tag it exactly. To solve this problem, we propose a hierarchical transformer with speaker embeddings (HiTransformer-s), which can model the interaction of speakers in a dialog.

For the $i-th$ dialog in D, $D_i = \{(u_j,\ s_j,\ c_j)\}_{j=1}^{N_i}$, the individual utterance embedding is $\{f(u_j)\}_{j=1}^{N_i}$ and $N_s(D_i)$ is the number of speakers in D_i. We first encode all the speakers in D_i with one-hot encoding and then pad them to the dimension of $Max\{N_s(D_i)\}_{i=1}^{L})$ with 0 to obtain the speaker embeddings $\{e_s(s_j)\}_{j=1}^{N_i}$.

$$\{e_s(s_j)\}_{j=1}^{N_i} = pad(onehot\Big(\{s_j\}_{j=1}^{N_i}\Big),\ Max\{N_s(D_j)\}_{j=1}^{L}) \tag{6}$$

Finally, we concatenate the summation of the individual utterance embeddings and the embeddings of position with the speaker embeddings of every utterance as the input of the upper-level transformer.

$$U = \{(f(u_j)\ \odot\ p_j) \oplus e_s(s_j)\}_{j=1}^{N_i} \tag{7}$$

where $\odot$ denotes element $-$ wise addition, and $\oplus$ is the concatenation operator.

3.4. *Model Training*

To solve the issue of class imbalance, following the above research [28], we use weighted cross entropy as the training loss to weight the samples of minority classes as below.

$$loss = -\frac{1}{\sum_{i=0}^{L} N_i} \sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{w_{c_j}} \sum_{c \in C} y_j^c log_2(\hat{y}_j^c) \tag{8}$$

$$w_c = \frac{a_c}{\sum_{i \in C} a_i} \tag{9}$$

where a_i denotes the number of utterances with emotion i in the training set.

4. Experimental Settings

In this section, we present the datasets, evaluation metrics, baselines, and experimental results of our model.

4.1. Dataset

Friends [29]: The dataset is annotated from the Friends TV Scripts, and each dialog in the dataset consists of a scene of multiple speakers. In total, there are 1000 dialogs, which are split into three parts: 720 for training, 80 for validation, and 200 dialogs for testing. Each utterance is tagged with an emotion in a set of emotions, {anger, joy, sadness, neutral, surprise, disgust, fear, and nonneutral}.

EmotionPush [29]: The dataset consists of private conversations between friends on Facebook and includes 1000 dialogs, which are split into 720, 80, and 200 dialogs for training, validation, and testing, respectively. Each utterance is tagged with an emotion in a set of emotions as in the Friends dataset.

EmoryNLP [19]: The dataset is annotated from the Friends TV Scripts as well. However, its size and annotations are different from the Friends dataset. It includes 713 dialogs for training, 99 dialogs for validation, and 85 dialogs for testing.

The emotion labels include neutral, sad, mad, scared, powerful, peaceful, and joyful. For the first two datasets, we follow a previous study [7] to consider only four emotion classes, i.e., anger, joy, sadness, and neutral, and consider all the emotion classes for EmoryNLP.

4.2. Evaluation Metrics

Following [4], who achieved the best performance on several ULER datasets, we choose the macro averaged F1-score as the primary metric for evaluating the performance of our models.

$$Macro - F1 = \frac{\sum_{c \in C} F1_c}{|C|} \tag{10}$$

where $F1_c$ is the F1-score of emotion c. We also report the weighted accuracy (WA) and unweighted accuracy (UWA), which were adopted in a previous study [30].

$$\text{WA} = \sum_{c \in C} w_c \cdot a_c \tag{11}$$

$$\text{UWA} = \frac{\sum_{c \in C} a_c}{|C|} \tag{12}$$

where w_c is the percentage of class c in the testing set, and a_c is the corresponding accuracy. As shown in Tables 1 and 2, most of the datasets in this paper have an imbalanced emotion distribution. Therefore, compared with WA and UWA, the F1-score is better for measuring the model performance.

Table 1. Detailed descriptions of Friends and EmotionPush.

Dataset	#Dialog (#Utterance)			Emotion				
	Train	Val	Test	Ang	Hap/Joy	Sad	Neu	Others
Friends	720 (10,561)	80 (1178)	200 (2764)	759	1710	498	6530	5006
EmotionPush	720 (10,733)	80 (1202)	200 (2807)	140	2100	514	9855	2133

Table 2. Detailed descriptions of EmoryNLP.

Dataset	#Dialog (#Utterance)			Emotion					
	Train	Val	Test	Neu	Joy	Peaceful	Powerful	Scared	Mad	Sad
EmoryNLP	713 (9934)	99 (1344)	85 (1328)	3776	2755	1191	1063	1645	1332	844

4.3. Compared Methods

We compare our model HiTransformer and HiTransformer-s with the following state-of-the-art baselines:

- SA-BiLSTM [21]: A self-attentive bidirectional LSTM model, an efficient model that achieved second place in the EmotionX Challenge [30];
- CNN-DCNN [28]: A convolutional-deconvolutional autoencoder with more handmade features, and the winner of the EmotionX Challenge [30];
- bcLSTM$_+$ [7]: A model with a 1-D CNN to extract the utterance embeddings and a bidirectional LSTM to model the relationship of utterances;
- bcGRU [7]: A variant of bcLSTM$_+$ with a BiGRU to capture the utterance-level context;
- CoDE$_{mid}$ [7]: CoDE$_{mid}$ is a context-dependent encoder (CoDE) model with a bidirectional GRU that extracts the utterance embeddings and a bidirectional GRU that models the relationship of utterances;
- $PT - CoDE_{mid}$ [7]: A variant of CoDE$_{mid}$ that pretrains a context-dependent encoder (CoDE) for ULER by learning from unlabeled conversation data;
- HiGRU [4]: A hierarchical gated recurrent unit (HiGRU) framework with a lower-level GRU to model the word-level inputs and an upper-level GRU to capture the contexts of utterance-level embeddings;
- HiGRU-f [4]: A variant of HiGRU with individual feature fusion;
- HiGRU-sf [4]: A variant of HiGRU with self-attention and feature fusion;
- SCNN [18]: A sequence-based convolutional neural network that utilizes the emotion sequence from the previous utterances for detecting the emotion of the current utterance.
- IDEA [8]: Two different BERT models were developed. For Friends, pretraining was done using a sliding window of two utterances to provide dialogue context. For EmotionPush, pretraining was performed on Twitter data, as it is similar in nature to chat-based dialogues. In both cases, special attention was given to the class imbalance issue by applying “weighted balanced warming” to the loss function.

4.4. Parameters

We adopt the pretrained uncased BERT-Base1 model as the lower-level transferable language model, where the maximum input length is 512. The number of combination layers of a multi-head attention and a feedforward neural network is 12. For the upper-level transformer layers, the number of transformer layers is 4, and the number of heads in the multi-head attention is 8. For the classification layer, the internal hidden size of the classification layer is set to 384, and the dropout rate is 0.5 to prevent overfitting. We adopt Adam [31] as the optimizer with a batch size of 1 and a learning rate of 1×10^{-5}. Early stopping with a patience of 5 is adopted to terminate training based on the accuracy of the validation set.

5. Result Analysis

We report the empirical results in Table 3, which show the average results of 10 trials each for the three datasets. From these results, we make the following observations.

Table 3. Testing results on Friend, EmotionPush, and EmoryNLP.

Model	**Friends**			**EmotionPush**			**EmoryNLP**		
	Mac-F1	**WA**	**UWA**	**Mac-F1**	**WA**	**UWA**	**Mac-F1**	**WA**	**UWA**
SA-BiLSTM	-	79.8	59.6	-	87.7	55.0	-	-	-
CNN-DCNN	-	67.0	62.5	-	75.7	62.5	-	-	-
bcLSTM$_+$	63.1	79.9	63.3	60.3	84.8	57.9	25.5	33.5	27.6
bcGRU	62.4	77.6	66.1	60.5	84.6	56.9	26.1	33.1	27.4
CoDE$_{mid}$	62.4	78.0	65.3	60.3	84.2	58.5	26.7	34.7	28.8
$PT - \text{CoDE}_{mid}$	65.9	81.3	66.8	62.6	84.7	60.4	29.1	36.1	30.3
HiGRU	-	74.4	67.2	-	73.8	66.3	-	-	-
HiGRU-f	-	71.3	68.4	-	73.0	66.9	-	-	-
HiGRU-sf	-	74.0	68.9	-	73.0	68.1	-	-	-
IDEA	**73.1**	**81.6**	**77.9**	**69.5**	**88.2**	**84.9**	-	-	-
SCNN	-	-	-	-	-	-	26.9	37.9	-
HiTransformer	66.66	82.11	63.71	63.90	86.87	61.55	31.36	37.25	29.24
HiTransformer-s	67.88	**82.18**	68.78	65.43	86.92	63.03	**33.04**	**37.98**	**32.67**

5.1. Comparison with Baselines

Our proposed HiTransformer-s obtains competitive results compared with the state-of-the-art methods in terms of the macro-F1 score. Specifically, HiTransformer-s obtains a 3.94% absolute improvement over EmoryNLP, and HiTransformer-s is 5.22% and 4.07% less than IDEA for Friends and EmorionPush, respectively. In addition, for Friends, HiTransformer-s obtains 0.58% improvement compared with the best performance in the past in terms of WA, and 9.12% less than the best performance from IDEA in terms of UWA. However, HiTransformer-s obtains a 0.58% improvement compared with IDEA in terms of WA. For EmotionPush, HiTransformer-s is 1.28% lower than IDEA in terms of WA and 21.87% lower than IDEA in terms of UWA. This can be attributed to different modifications of model IDEA based on different data sets, and pretraining was performed on Twitter data. IDEA includes two different BERT models to deal with two different datasets, and pretraining was performed with different data sets. Although IDEA has done well in the above two datasets, the method has not done well in terms of generality. For EmoryNLP, HiTransformer-s obtains 1.88% and 2.37% absolute improvement in terms of WA and UWA, respectively. To make the results more intuitive and prominent, in Figure 5a, we compare the $Macro - F1$ values of several excellent methods for the three datasets. The above results demonstrate the superior power of HiTransformer-s and HiTransformer.

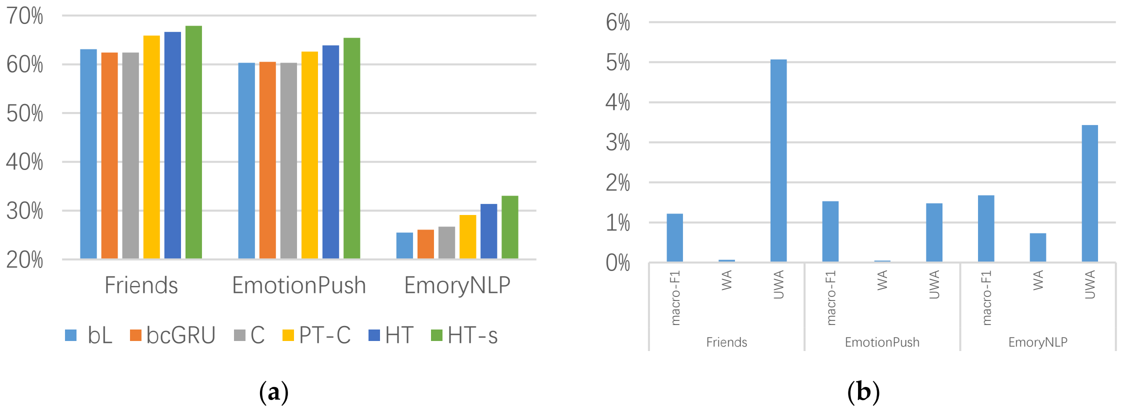

Figure 5. (**a**) The $Macro - F1$ of several excellent methods. bL: bcLSTM$_+$, bG: bcGRU, C:CoDE$_{mid}$,PT-C: $PT - \text{CoDE}_{mid}$, HT: HiTransformer, HT-s: HiTransformer-s; (**b**) Description of the difference between the results of HiTransformer-s and HiTransformer applied to three datasets.

5.2. HiTransformer vs. HiTransformer-s

By analyzing ULER, we find that speaker information plays an important role in utterance classification. Therefore, we proposed HiTransformer-s on the basis of HiTransformer. From Table 3, we observe that HiTransformer-s outperforms HiTransformer for all three datasets in terms of macro-F1, WA, and UWA. Specifically, for Friends, HiTransformer-s attains 1.22%, 0.07%, and 5.07% improvement over HiTransformer in terms of macro-F1, WA, and UWA, respectively. For EmotionPush, HiTransformer-s attains 1.53%, 0.05%, and 1.48% improvement over HiTransformer in terms of macro-F1, WA, and UWA, respectively. For EmoryNLP, HiTransformer-s attains 1.68%, 0.73%, and 3.43% improvement over HiTransformer in terms of macro-F1, WA, and UWA, respectively. To make the results more intuitive and prominent, in Figure 5b, we compare the results of HiTransformer and HiTransformer-s for the three datasets. The vertical axis is the difference between the results of HiTransformer-s and HiTransformer. The results demonstrate that HiTransformer-s with speaker embedding outperforms HiTransformer for all three datasets Friends, EmotionPush, and EmoryNLP, which also proves that speaker information is necessary to deal with the problem of ULER. Finally, we test the training and testing speed of HiTransformer-s and HiTransformer using EmoryNLP. The training and testing times of HiTransformer are 83.4 s and 4.8 s, respectively, and those of HiTransformer-s are 78.3 s and 5.2 s, respectively. The speed of the two models is similar.

5.3. Error Analysis

Figure 6 shows the confusion matrix of gold labels and the prediction for our best model, HiTransformer-s, for dataset Friends. We show only four types of emotions that we care about. Mostly, all of the emotions get confused the most with Neu. The utterances that belong to Ang and Sad are more likely to be confused than those that belong to Ang and Joy or Sad and Joy. This may be due to *Ang* and Sad expressing more similar emotions. How to differentiate intrinsically similar emotions and how to differentiate Neu and other emotions are two challenging directions in this field.

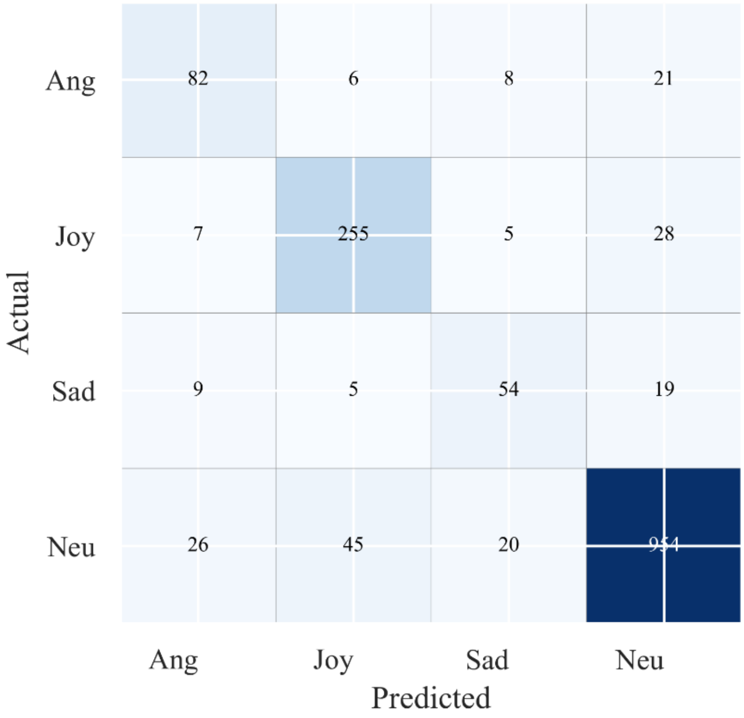

Figure 6. Confusion matrix of HiTransformer-s for the evaluation set of Friend.

6. Conclusions

In this work, to address utterance-level emotion recognition in dialog systems, we propose a hierarchical transformer (HiTransformer) framework with a lower-level transformer to model word-level input and an upper-level transformer to capture the contexts of utterance-level embeddings. To obtain better individual utterance embeddings, we adopt BERT, which is pretrained on a massive unlabeled corpus as the lower-level transformer. To enable HiTransformer to obtain speaker information, we propose HiTransformer-s. Experimental results demonstrate that our proposed hierarchical transformer models can sufficiently capture the available text information.

In the future, we plan to pretrain a transformer model to capture the relationship of utterances, similar to BERT, and adopt it as the upper-level transformer to capture the contextual information more sufficiently, which can also address the problem of data scarcity in ULER. In addition, in the real-time dialog systems, our models have no information regarding the future. Thus, the performance of the approach is expected to decrease. It would be interesting to deal with this issue.

Author Contributions: Conceptualization, Q.L. and C.W.; Data curation, Q.L.; Methodology, Z.W. and Q.L.; Validation, Z.W.; Writing—original draft, Q.L.; Writing—review & editing, Z.W.; Project administration, C.W. and K.Z.; funding acquisition, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the National Key R&D Program of China under Grant No. 2017YFB0802703 and the Natural Science Foundation of China under Grant No. 61602052.

Acknowledgments: We also thank my graduate tutor ChunHua Wu for its linguistic assistance during the revision of this manuscript, and she also provided the servers used for experiments. In addition, we also thank the three anonymous reviewers for the insightful suggestions on various aspects of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883.
2. Olson, D. From utterance to text: The bias of language in speech and writing. *Harv. Educ. Rev.* **1977**, *47*, 257–281. [CrossRef]
3. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.; Zimmermann, R. Conversational memory network for emotion recognition in dyadic dialogue videos. In Proceedings of the North American Chapter of the Association for Computational Linguistics, New Orleans, LA, USA, 1–6 June 2018; pp. 2122–2132.
4. Jiao, W.; Yang, H.; King, I.; Lyu, M.R. HiGRU: Hierarchical Gated Recurrent Units for Utterance-level Emotion Recognition. In Proceedings of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 397–406.
5. Bradbury, J.; Merity, S.; Xiong, C.; Socher, R. Quasi-recurrent neural networks. *arXiv* **2016**, arXiv:1611.01576.
6. Zhong, P.; Wang, D.; Miao, C. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In Proceedings of the International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 165–176.
7. Jiao, W.; Michae, R.L.; Irwin, K. PT-CoDE: Pre-trained Context-Dependent Encoder for Utterance-level Emotion Recognition. *arXiv* **2019**, arXiv:1910.08916.
8. Huang, Y.; Lee, S.; Ma, M.; Chen, Y.; Yu, Y.; Chen, Y. EmotionX-IDEA: Emotion BERT—An Affectional Model for Conversation. *arXiv* **2019**, arXiv:1908.06264.
9. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
11. Navarretta, C.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B. Mirroring facial expressions and emotions in dyadic conversations. In Proceedings of the Language Resources and Evaluation Conference, Portoroz, Slovenia, 23–28 May 2016; pp. 469–474.

12. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998; pp. 137–142.
13. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2013**, *3*, 1137–1155.
14. Socher, R.; Pennington, J.; Huang, E.H.; Ng, A.Y.; Manning, C.D. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 151–161.
15. Kim, Y. Convolutional neural networks for sentence classification. In Proceedings of the Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1746–1751.
16. Abdul-Mageed, M.; Ungar, L.H. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In Proceedings of the Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 718–728.
17. Poria, S.; Cambria, E.; Gelbukh, A.F. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2539–2544.
18. Luo, L.; Yang, H.; Chin, F.Y.L. Emotionx-dlc: Self-attentive BiLSTM for detecting sequential emotions in dialogues. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Melbourne, Australia, 20 July 2018; pp. 32–36.
19. Zahiri, S.M.; Choi, J.D. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 44–52.
20. Zhou, X.; Li, L.; Dong, D.; Liu, Y.; Chen, Y.; Zhao, W.X.; Yu, D.; Wu, H. Multi-turn response selection for chatbots with deep attention matching network. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1118–1127.
21. Dai, Z.; Yang, Z.; Yang, Y.; Cohen, W.W.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2978–2988.
22. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/languageunderstandingpaper.pdf (accessed on 23 June 2020).
23. Dehghani, M.; Gouws, S.; Vinyals, O.; Uszkoreit, J.; Kaiser, Ł. Universal transformers. *arXiv* **2018**, arXiv:1807.03819.
24. Liu, P.; Chang, S.; Huang, X.; Tang, J.; Cheung, J.C.K. Contextualized non-local neural networks for sqeunce learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 6762–6769.
25. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, A.L. Deep contextualized word representations. In Proceedings of the North American Chapter of the Association for Computational Linguistics, New Orleans, LA, USA, 1 June 2018; pp. 2227–2237.
26. Sun, C.; Huang, L.; Qiu, X. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In Proceedings of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 380–385.
27. Nils, R.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 3973–3983.
28. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
29. Hsu, C.C.; Ku, L.W. Socialnlp 2018 emotionx challenge overview: Recognizing emotions in dialogues. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Melbourne, Australia, 20 July 2018; pp. 27–31.

30. Khosla, S. Emotionx-ar: CNN-DCNN autoencoder based emotion classifier. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Melbourne, Australia, 20 July 2018; pp. 37–44.
31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.