

Conversational Transfer Learning for Emotion Recognition

Devamanyu Hazarika^a, Soujanya Poria^{c,*}, Roger Zimmermann^a, Rada Mihalcea^b

^a*School of Computing, National University of Singapore*

^b*Computer Science & Engineering, University of Michigan, USA*

^c*Information Systems Technology and Design, Singapore University of Technology and Design*

Abstract

Recognizing emotions in conversations is a challenging task due to the presence of contextual dependencies governed by self- and inter-personal influences. Recent approaches have focused on modeling these dependencies primarily via supervised learning. However, purely supervised strategies demand large amounts of annotated data, which is lacking in most of the available corpora in this task. To tackle this challenge, we look at transfer learning approaches as a viable alternative. Given the large amount of available conversational data, we investigate whether generative conversational models can be leveraged to transfer affective knowledge for detecting emotions in context. We propose an approach, *TL-ERC*, where we pre-train a hierarchical dialogue model on multi-turn conversations (source) and then transfer its parameters to a conversational emotion classifier (target). In addition to the popular practice of using pre-trained sentence encoders, our approach also incorporates recurrent parameters that model inter-sentential context across the whole conversation. Based on this idea, we perform several experiments across multiple datasets and find improvement in performance and robustness against limited training data. TL-ERC also achieves better validation performances in significantly fewer epochs. Overall, we infer that knowledge acquired from dialogue generators can indeed help recognize emotions in conversations.

Keywords: Emotion Recognition in Conversations, Transfer Learning, Generative Pre-training, Conversation Modeling

1. Introduction

Emotion Recognition in Conversations (ERC) is the task of detecting emotions from utterances in a conversation. It is an important task with applications

*Corresponding author.

Email addresses: hazarika@comp.nus.edu.sg (Devamanyu Hazarika), soujanya_poria@sutd.edu.sg (Soujanya Poria), rogerz@comp.nus.edu.sg (Roger Zimmermann), mihalcea@umich.edu (Rada Mihalcea)

ranging from dialogue understanding to affective dialogue systems [1]. Apart from the traditional challenges of dialogue understanding, such as intent-detection, contextual grounding, and others [2], ERC presents additional challenges as it requires the ability to model emotional dynamics governed by self- and inter-speaker influences at play [3]. Further complications arise due to the limited availability of annotated data — especially in multimodal ERC — and the variability in annotations owing to the subjectivity of annotators in interpreting emotions.

In this work, we focus on these issues by investigating a framework of sequential inductive *transfer learning* (TL) [4]. In particular, we attempt to transfer contextual affective information from a generative conversation modeling task to ERC. We name this framework TL-ERC.

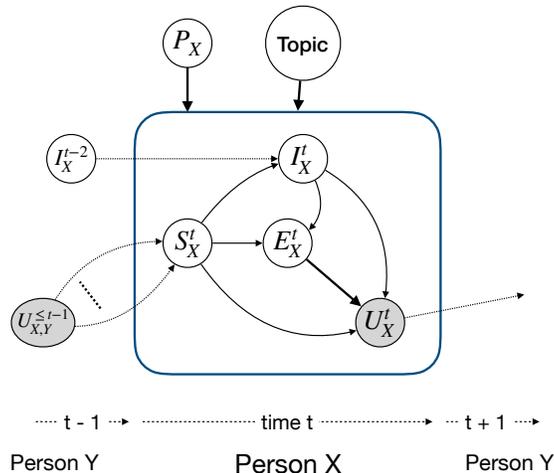


Figure 1: Dyadic conversation — between person X and Y — are governed by interactions between several latent factors. Emotions are a crucial component in this generative process. In the illustration, P represents the personality of the speaker; S represents speaker-state; I denotes the intent of the speaker; E refers to the speaker’s emotional state, and U refers to the observed utterance. Speaker personality and the topic always condition upon the variables. At turn t , the speaker conceives several pragmatic concepts such as argumentation logic, viewpoint, and inter-personal relationship - which we collectively represent using the speaker-state S [5]. Next, the intent I of the speaker gets formulated based on the current speaker-state and previous intent of the same speaker (at $t-2$). These two factors influence the emotional feeling of the speaker, which finally manifests as the spoken utterance [1].

But why should generative modeling of conversations acquire knowledge on emotional dynamics? To answer this question, we first observe the role of emotions in conversations. Several works in the literature have indicated that emotional goals and influences act as latent controllers in dialogues [6, 7]. Poria et al. [1] demonstrated the interplay of several factors, such as the topic of the conversation, speakers’ personality, argumentation-logic, viewpoint, and intent, which modulate the emotional state of the speaker and finally lead to an utterance. Fig. 1 illustrates these dependencies, which elaborate emotional

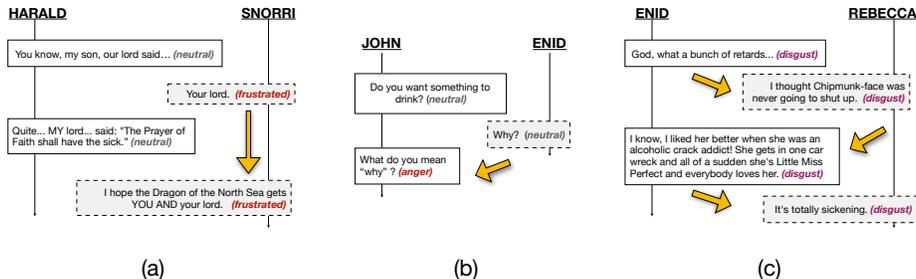


Figure 2: Samples from *Cornell Movie Dialog Corpus* [11]. The examples demonstrate various kinds of emotional influences, such as *emotional inertia*, *mirroring*, etc., that manifest in natural conversations.

factor as a critical latent state in the overall generative process of dialogues.

The interactions between these latent factors lead to diverse emotional dynamics within the conversations. Fig. 2 provides some examples demonstrating such patterns. In the figure, conversation (a) illustrates the presence of *emotional inertia* [8] which occurs though self-influences in emotional states. The character *Snorri* maintains a frustrated emotional state by not being affected/influenced by the other speaker. Whereas, conversation (b) and (c) demonstrate the role of inter-speaker influences in emotional transitions across turns. In (b), the character *John* is triggered for an *emotional shift* due to influences based on his counterpart’s responses, while (c) demonstrates the effect of *mirroring* [9] which often arises due to topical agreement between speakers. All these examples demonstrate the presence of such emotional dynamics that are not just inherent in the conversations but also help shape them up [1].

To model such conversations, a generator would require the ability to 1) interpret latent emotions from its contextual turns and 2) model the complex dynamics governing them. In addition, it would also need to interpret other factors such as topic of the conversations, speaker personalities, intents, etc. Such a model would then be a *perfect* dialogue generator. We illustrate this in Fig. 3, where the model generating utterance utt_{t+1} would require to understand the emotions of the context arising from the utterances utt_t, utt_{t-1} , and so on. Thereby, we hypothesize that a trained dialogue generator would possess the ability to model implicit affective patterns across a conversation [10]. Consequently, we propose a framework that uses TL to transfer this affective knowledge into our target discriminative task, i.e., ERC.

In our approach, we first pre-train a hierarchical generative dialogue model on the *source task* of conversation modeling. Being an unsupervised (or self-supervised) task, conversation modeling typically benefits from a large amount of data in the form of multi-turn chats. Next, we *adapt* our model to the *target task* (ERC) by transferring the inter-sentence contextual parameters ¹ from the

¹In this paper, *context* refers to the inter-sentential context in conversations, i.e. the

trained source model. For sentence encoding, we choose the BERT model [12], which is pre-trained on masked language modeling and next sentence prediction objectives.

Although we acknowledge that training a *perfect* dialogue generator is presently challenging, we demonstrate that benefits can be observed even with a popular baseline generator. In the bigger picture, our approach can enable the co-evolution of both generative and discriminative models for the tasks mentioned above. This is possible since improving an emotional classifier using a dialogue model can, in turn, be utilized to enhance dialogue models with emotional intelligence further, leading to an iterative cycle of improvements for both the applications.

Overall, our contributions are summarized as follows:

- We propose *TL-ERC*, which pre-trains a hierarchical generative dialogue model on multi-turn conversations (*source*) and subsequently transfers affective knowledge to the *target* task of ERC. Despite the active role of TL in providing state-of-the-art token and sentence encoders, its use in leveraging multi-turn contextual knowledge — across utterances — has been mostly unexplored. Our work stands as one of the first in this direction.
- Through our experiments, we observe the promising effects of using these pre-trained weights. Our models, initialized with the acquired knowledge, converge faster compared to randomly initialized counterparts and also demonstrate robust performance in limited training-data scenarios.
- We identify various challenges observed in using TL for ERC. These points raise essential research questions and provide a roadmap for future research in this topic.

In the remaining paper, Section 2 first discusses the works in the literature related to our task and our approach. Next, Section 3 provides information on the TL setup along with details on the design of the framework. Experimental details are mentioned in Section 4; results and extensive analyses are provided in Section 5. Section 6 provides some challenges observed in the proposed framework, casting light for future research efforts. Finally, Section 7 concludes the paper.

2. Related Works

We proceed to discuss the use of transfer learning by the available literature in NLP. First, we enlist some of the famous works that have benefited from TL, and then we focus on works that attempt to frame TL in the context-level hierarchy. Next, we look at recent works on emotion/sentiment analysis, including works that have employed TL. Finally, we attempt to position our contribution amidst the latest developments in ERC.

sequential information acquired from utterances of speakers in a conversation.

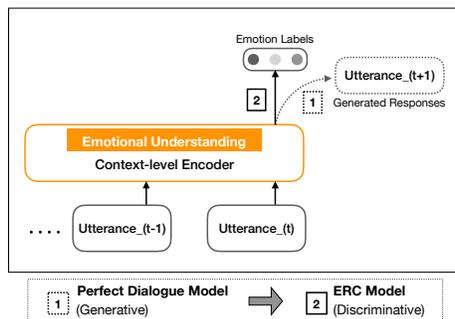


Figure 3: The figure illustrates how a perfect dialogue generator requires emotional understanding from its context – a transferrable knowledge into ERC.

2.1. Transfer Learning in NLP

Transfer learning has played a critical role in the success of modern-day NLP systems. As a matter of fact, the key milestones in the recent history of NLP are provided by works using TL. NLP has particularly benefited by inductive TL, where unlabelled data is utilized to leverage knowledge for labeled downstream tasks. Early works, such as Ando and Zhang [13], introduced this concept, which was heavily adopted by the community and has shown tremendous success ever since [14].

Modern breakthroughs, such as neural word embeddings, followed similar modeling by utilizing unlabeled textual data to learn the embeddings [15]. Of late, there has been a significant interest in using language models (LMs) to learn contextual embeddings [16, 17]. TL through LM pre-training has also provided state-of-the-art text classifiers with high quality sentence encoders [18, 12, 19]. Consequently, several works have explored improving this framework by either modifying the LM pre-training approach or weight adaptation in the downstream tasks [20, 21].

Context-level Transfer Learning. Availability for works that explore TL for inter-sentence or sequential learning is limited. Some of these include sentence-level sequence tagging tasks [22] or inter-sentence supervised tasks such as query matching in conversations [23], next sentence prediction [12], etc. Recent works that address the topic of pre-training sentence representations or multi-turn conversations follow either a retrieval-based or a generative strategy. For the former, strategies include contrastive sentence selection (ToD-BERT [24], ConveRT [25]), sentence ordering (ALBERT [26]), and semantical sentence matching (Sentence-BERT [27]) objectives. Whereas, generative models attempt to learn a probabilistic model for the conversations directly. DialoGPT [28] is a recently proposed model that proposes a generative model based on the GPT architecture [29]. Our pre-training model is similar in spirit to DialoGPT. However, we do not flatten the conversation and instead opt for a hierarchical conversation model. This also suits our downstream task of conversational

emotion recognition. Additionally, we analyze the joint pre-training of full conversations in a self-supervised setting and attempt to observe its efficacy in transferring affective knowledge.

2.2. Affect Analysis

Affect, in particular emotions, are an integral part of human life and modulate our day-to-day behavior and activities [30]. The interest in understanding emotions is multi-disciplinary and covers a long history of research. The importance of modeling emotions has multiple benefits across applications such as e-learning [31], human-computer interaction [32], user profiling [33], etc.

From a computational perspective, emotions are typically studied across various media formats, covering applications such as facial emotion recognition [34, 35], emotions in speech [36, 37], or multimodal emotion recognition [38]. In text-based applications, machine learning has played a crucial role in mining emotions [39]. Earlier approaches designed hand-crafted features that included emotional keyword spotting [40], affect-based lexical resources (WordNet-Affect [41], SentiWordNet [42]), and distant supervision via hashtags [43]. In the present deep-learning era,

In the present deep learning era, approaches have diverged from hand-crafted features and moved towards automated feature learning. Modern approaches consider advanced neural architectures, such as convolutional networks [44], recurrent networks [45], and attention mechanisms [46] for emotion detection. Recent times have also seen approaches that address practical scenarios such as domain awareness [47], and utilize alternate training strategies, such as adversarial approaches [48]. Complementary to these issues, we address data scarcity issues in ERC and leverage transfer learning for the same. We discuss the related works aligned to these topics next.

Transfer Learning for Affect. TL for affective analysis has gained momentum in recent years, with several works adopting TL-based approaches for their respective tasks. These works leverage diverse source tasks, such as, sentiment/emotion analysis in text [49, 50, 51], large-scale image classification in vision [52], sparse auto-encoding in speech [53], etc. Felbo et al. utilize emojis present in online platforms to pre-train models and transfer knowledge for emotion recognition. Using layer-wise fine-tuning, they also transfer knowledge into related tasks of sarcasm and sentiment detection. A similar approach is taken by Daval-Frerot et al.. Similar to these works, our approach also leverages TL for knowledge transfer. However, our task is in a sequential setting at the conversational level. To the best of our knowledge, our work is one of the first that explores TL in ERC and utilizes generative conversation modeling as a pre-training objective.

Emotion Recognition in Conversations. ERC is an emerging sub-field of affective computing and is developing into an active area of research. Current works try to model contextual relationships amongst utterances in a supervised fashion to model the implicit emotional dynamics. Strategies include modeling speaker-based dependencies using recurrent neural networks [55, 56], memory networks [3,

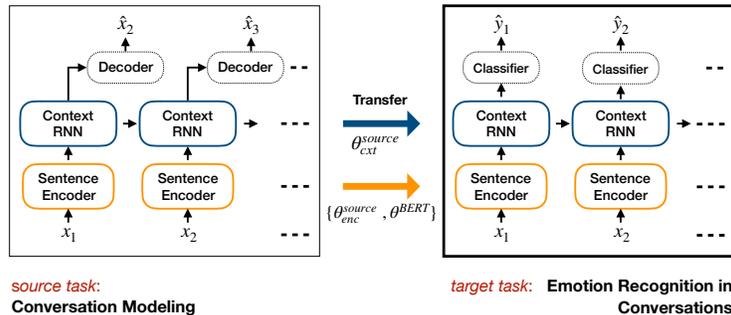


Figure 4: Proposed framework for ERC using TL parameters. The model on the left is a conversational response generator which is used as a pre-trained model. The parameters are transferred to the target model as shown on the right side.

57], graph neural networks [58, 59], quantum-inspired networks [60], amongst others. Some of these works also explore challenges such as multi-speaker modeling [61], multimodal processing [57], and knowledge infusion [62]. BERT-based sentence encoding has also been heavily adopted by the latest works in this area [63]. Works like EmotionX-IDEA [64] and PT-Code [65], developed concurrently to ours, follow a similar vein by transferring emotional knowledge from BERT pre-training. However, in these works, either the conversations are limited to utterance-reply pairs or follow a contrastive utterance retrieval objective. Our work, in contrast, pre-trains a whole conversation jointly using a hierarchical generative model. Overall, we find that there is a dearth of works that consider scarcity issues for annotated data and leverage TL. Our work strives to fill this gap by providing a systematic study for TL in ERC.

3. Methodology

Our proposed framework, TL-ERC, is summarized in Fig. 4. First, we define the source generative model trained as a dialogue generator, followed by a description of the target model, which performs hierarchical context encoding — for the task of ERC — using BERT-based sentence encoders and learned context weights from the source model.

3.1. Source: Generative Conversation Modeling

To perform the generative task of conversation modeling, we use the *Hierarchical Recurrent Encoder-Decoder* (HRED) architecture [66]. HRED is a classic framework for seq2seq conversational response generation that models conversations in a hierarchical fashion using three sequential components: *encoder* recurrent neural networks (RNNs) for sentence encoding, *context* RNNs for modeling the conversational context across sentences, and *decoder* RNNs for generating the response sentence.

For a given conversation context with sentences x_1, \dots, x_t , HRED generates the response x_{t+1} as follows:

1. **Sentence Encoder:** It encodes each sentence in the context using an *encoder RNN*, such that,

$$\mathbf{h}_t^{enc} = f_{\theta}^{enc}(\mathbf{x}_t, \mathbf{h}_{t-1}^{enc})$$

2. **Context Encoder:** The sentence representations are then fed into a *context RNN* that models the conversational context until time step t as

$$\mathbf{h}_t^{cxt} = f_{\theta}^{cxt}(\mathbf{h}_t^{enc}, \mathbf{h}_{t-1}^{cxt})$$

3. **Sentence Decoder:** Finally, an auto-regressive *decoder RNN* generates sentence x_{t+1} conditioned on \mathbf{h}_t^{cxt} , i.e.,

$$\begin{aligned} p_{\theta}(x_{t+1}|x_{\leq t}) &= f_{\theta}^{dec}(x | \mathbf{h}_t^{cxt}) \\ &= \prod_i f_{\theta}^{dec}(x_{t+1,i} | \mathbf{h}_t^{cxt}, x_{t+1,<i}) \end{aligned}$$

With the i^{th} conversation being a sequence of utterances $C_i = [x_{i,1}, \dots, x_{i,n_i}]$, HRED trains all the conversations in the dataset together by using the maximum likelihood estimation objective $\arg \max_{\theta} = \sum_i \log p_{\theta}(C_i)$.

The HRED model provides the possibility to introduce multiple complexities in the form of multi-layer RNNs and other novel encoding strategies. In this work, we choose to experiment with the original version of the architecture with single-layer components so that we can analyze the hypothesis without unwanted contribution from the added complexities. In our source model, f_{θ}^{enc} can be any RNN function, which we model using the bi-directional *Gated Recurrent Unit* (GRU) variant [67] to encode each sentence. We call the parameters associated with this GRU function as θ_{enc}^{source} . For both the *context RNN* (f_{θ}^{cxt}) and *decoder RNN*, we use uni-directional GRUs — with parameters θ_{cxt}^{source} and θ_{dec}^{source} , respectively — and complement the decoder with beam-decoding for generation ².

3.2. Target: Emotion Recognition in Conversations

The input for this task is also a conversation C with constituent utterances $[x_1, \dots, x_n]$. Each x_i is associated with an emotion label $y_i \in \mathbb{Y}$. We adopt a setup similar to the three components described for the source task, as in Poria et al. [68]. However, the *decoder* in this setup is replaced by a discriminative mapping to the label space instead of a generative network. Below, we describe the different initialization parameters that we consider for the first two stages of the network:

²Model implementations are adapted from <https://github.com/ctr4si/>

3.2.1. Sentence Encoding

To encode each utterance in the conversation, we consider the state-of-the-art universal sentence encoder BERT [12], with its parameters represented as θ^{BERT} . We choose BERT over the HRED sentence encoder (θ_{enc}^{source}) as it provides better performance (see Table 8). Also, BERT includes the task of next sentence prediction as one of its training objectives which aligns with the inter-sentence level of abstraction that we consider in this work.

We choose the *BERT-base uncased* pre-trained model as our sentence encoder³. Though this model contains 12 transformer layers, to limit the total number of parameters in our model, we restrict to the first 4 transformer layers. To get a sentential representation, we use the hidden vectors of the first token [CLS] across the considered transformer layers (see Devlin et al. [12]) and mean-pool them to get the final sentence representation.

3.2.2. Context Encoding

We use a similar context encoder RNN as the source HRED model with the option to transfer the learned parameters θ_{cxt}^{source} . For input sentence representation \mathbf{h}_t^{enc} provided by the *encoder RNN*, the *context RNN* transforms it as follows:

$$\begin{aligned} \mathbf{z}_t &= \sigma(V^z \mathbf{h}_t^{enc} + W^z \mathbf{h}_{t-1}^{cxt} + \mathbf{b}^z) \\ \mathbf{r}_t &= \sigma(V^r \mathbf{h}_t^{enc} + W^r \mathbf{h}_{t-1}^{cxt} + \mathbf{b}^r) \\ \mathbf{v}_t &= \tanh(V^h \mathbf{h}_t^{enc} + W^h (\mathbf{h}_{t-1}^{cxt} \otimes \mathbf{r}_t) + \mathbf{b}^h) \\ \mathbf{h}_t^{cxt} &= (1 - \mathbf{z}_t) \otimes \mathbf{v}_t + \mathbf{z}_t \otimes \mathbf{h}_{t-1}^{cxt} \\ \mathbf{h}_t^{cxt} &= \tanh(W^p \mathbf{h}_t^{cxt} + \mathbf{b}^p) \end{aligned}$$

Here, $\{V^{z,r,h}, W^{z,r,h}, \mathbf{b}^{z,r,h}\}$ are parameters for the GRU function and $\{W^p, \mathbf{b}^p\}$ are additional parameters of a dense layer. For our setup, adhering to size considerations, we consider our transfer parameters to be $\theta_{cxt}^{source} = \{W^{z,r,h,p}, \mathbf{b}^{z,r,h,p}\}$.

3.2.3. Classification

For each turn in the conversation, the output from the context RNN is projected to the label-space, which provides the predicted emotion for the associated utterance. Similar to HRED, we train for all the utterances in the conversation together using the standard *Cross Entropy* loss. For regression targets, we utilize the *Mean Square Error (MSE)* loss, instead.

4. Experimental Setup

In this section, we define the experimental setup followed in this work. First, we detail the datasets that we utilize and mention their properties. Further, we provide information on the metrics used for evaluation, the training setup, and the model variants considered to test our hypothesis.

³<https://github.com/huggingface/pytorch-pretrained-BERT>

4.1. Datasets

4.1.1. Source Task

For pre-training with the source task of conversation modeling, we consider two large-scale benchmark datasets:

- *Cornell Movie Dialog Corpus* [11] is a popular collection of fictional conversations extracted from movie scripts. In this dataset, conversations are sampled from a diverse set of 617 movies leading to over 83k dialogues.
- *Ubuntu Dialog Corpus* [69] is a larger corpus with around 1 million dialogues, which, like the Cornell corpus, comprises of unstructured multi-turn dialogues based on Ubuntu chat logs (Internet Relay Chat).

Both datasets contain dyadic, i.e. two-party conversations. For brevity, throughout the paper, we mention these datasets as Cornell and Ubuntu, respectively. The data splits for training are created as per Park et al. [70].

4.1.2. Target Task

For the target task of ERC, we experiment with three datasets popular in this area of research:

- Primarily, we consider the textual modality of a small-sized multimodal dataset *IEMOCAP* [71] consisting of dyadic conversations between 10 speakers. Each pair is assigned one of many diverse conversational scenarios, with a total of five sessions across the dataset. Each conversational video is segmented into utterances and annotated with the following emotion labels: *anger*, *happiness*, *sadness*, *neutral*, *excitement*, and *frustration*. Split creating scheme is based on Hazarika et al. [57].
- We also analyze results on a moderately-sized emotional dialogue dataset *DailyDialog* [72] with labeled emotions: *anger*, *happiness*, *sadness*, *surprise*, *fear*, *disgust* and *no_emotion*. Unlike spoken utterances in IEMOCAP, the conversations are chat-based based on daily life topics. For creating the splits, we follow the original split details provided by Li et al. [72].
- Finally, we choose a regression-based dataset *SEMAINE*, which is a video-based corpus of human-agent emotional interactions. We use the split configuration detailed in AVEC 2012’s *fully continuous sub-challenge* [73] for the prediction of affective dimensions: *valence*, *arousal*, *power*, and *expectancy*. Annotation configuration is based on Hazarika et al. [57].

Table 1 provides the sizes along with split distributions for the above-mentioned datasets. For both IEMOCAP and SEMAINE, we generate the validation sets by random-sampling of 20% dialogue videos from the training sets. The class distribution for the categorical emotions in IEMOCAP and DailyDialog are presented in Table 2. From the table, IEMOCAP is observed a fairly balanced dataset whereas DailyDialog is highly skewed towards sentences with no emotion. As such, we decide upon different metrics for each dataset as discussed next.

Dataset		Dataset splits			
		train	validation	test	
Source	Cornell	#D	66,477	8,310	8,310
		#U	244,030	30,436	30,247
	Ubuntu	#D	898,142	18,920	19,560
		#U	6,893,060	135,747	139,775
Target	IEMOCAP	#D	120	31	
		#U	5810	1,623	
	SEMAINE	#D	58	22	
		#U	4386	1,430	
	Dailydialog	#D	11,118	1,000	1,000
		#U	87,170	7,740	8,069

Table 1: Table illustrates the sizes of the datasets used in this work. #D represents the number of dialogues whereas #U represents the total number of constituting utterances.

	Iemocap		Dailydialog		
	train/val	test	train	val	test
hap	504	144	11182	684	1019
sad	839	245	969	79	102
neu	1324	384	72143	7108	6321
ang	933	170	827	77	118
exc	742	299	-	-	-
frus	1468	381	-	-	-
surp	-	-	1600	107	116
fear	-	-	146	11	17
disg	-	-	303	3	47

Table 2: Category-wise distribution of utterances. *hap*: happiness; *neu*: neutral or no emotion; *ang*: anger; *exc*: excitement; *frus*: frustration; *surp*: surprise; *disg*: disgust.

4.1.3. Metrics

We choose the pre-training weights from the source task based on the best validation perplexity score [70]. For ERC, we use *weighted-F-score* metric for the classification tasks on IEMOCAP and DailyDialog. For DailyDialog, we remove *no_emotion* class from the F-score calculations due to its high majority (82.6%/81.3% occupancy in training/testing set) which hinders evaluation of other classes⁴. For the regression task on SEMAINE, we take the Pearson correlation coefficient (r) as its metric.

We also provide the average *best epoch* (BE) on which the least validation losses — across the multiple runs — are observed, and the testing evaluations are performed. A lower BE represents the model’s ability to reach optimum performance in lesser training epochs.

4.2. Model Size

We consider two versions of the source generative model: **HRED-small** and **HRED-large** with 256 and 1000-dimensional hidden state sizes, respectively. While testing the performance of both the models on the IEMOCAP dataset, we find the context weights from *HRED-small* (Cornell dataset) to provide better performance on average (58.5% F-score) over *HRED-large* (55.3% F-score). Following this observation, and also to avoid over-fitting on the small target datasets due to increased parameters, we choose the *HRED-small* model as the source task model for our TL procedure.

4.3. Model Variants and Baselines

The primary goal of this paper is to analyze the effect of TL at the conversation level for ERC. For this, we experiment on different variants of our model based on the parameter initialization procedure. We provide a summary of these variants in Table 3. In the table, *Variant 1* is the model with randomly initialized

⁴Evaluation strategy adapted from Semeval 2019 ERC task: www.humanizing-ai.com/emocontext.html

Variant	Initial Weight		Model Description
	sent _{enc}	cxt _{enc}	
(1)	-	-	Sentence encoders – <i>randomly</i> initialized. Context encoders – <i>randomly</i> initialized.
(2)	θ^{BERT}	-	Sentence encoders – BERT parameters. Context encoders – <i>randomly</i> initialized.
(3)	θ^{BERT}	$\theta^{ubuntu/cornell}_{cxt}$	TL-ERC Sentence encoders – BERT parameters. Context encoders – initialized from generative models pre-trained on Ubuntu/Cornell corpus.

Table 3: Variants of the model used in the experiments. Variant (3) is the proposed TL-ERC model.

parameters. In *Variant 2*, we replace the sentence encoders with the BERT model including its original pre-trained parameters. Finally, in *Variant 3*, in addition to BERT sentence encoders, we also initialize the context-RNN parameters learned from the source task. Different results and analyses amongst these variants are provided in Section 5.

Next, to compare our model with the existing literature, we select some prior state-of-the-art models evaluated on the target datasets:

- CNN [74] extracts textual features based on Convolutional Neural Networks (CNN). This is a non-contextual model, which evaluates each utterance in a conversation independently.
- Memnet [75] assigns dedicated memory for each historical utterance and performs multi-hop inference on them to get final representations for emotion classification.
- c-LSTM [68] is a popular model which is similar to our target model. It employs a bi-directional LSTM [76] to capture inter-utterance dependencies.
- c-LSTM+Att [77] enhances the c-LSTM model with inter-modality and inter-utterance attention mechanisms.
- CMN [3], the Conversational Memory Network, is an extension to the Memnet model which allots separate memories to both speakers in a dyadic conversational exchange.
- DialogueRNN [61] is a strong state-of-the-art baseline which employs three stages of recurrent units comprising global, speaker-state, and emotional units. The global RNN models the conversational context, speaker-state RNN models the individual speaker-states, and emotion RNN models the final emotional representations used for classification. For a fair comparison with our model, we chose the basic version of DialogueRNN without bi-directional RNNs and inter-utterance attention mechanisms.

Results on these baselines are provided in Section 5.5.

Variant	Initial Weights		Dataset: IEMOCAP							
			10%		25%		50%		100%	
	sent _{enc}	cxt _{enc}	F-Score	BE	F-Score	BE	F-Score	BE	F-Score	BE
(1)	-	-	23.2 ±0.4	48.4	41.6 ±0.8	72.5	48.4 ±0.3	75.1	53.8 ±0.3	13.8
(2)	θ^{BERT}	-	32.4 ±1.1	11.0	41.9 ±0.5	8.0	49.2 ±1.0	6.3	55.1 ±0.6	5.0
(3)	θ^{BERT}	θ_{cxt}^{ubuntu} $\theta_{cxt}^{cornell}$	35.7 ±1.1	14.2	45.9 ±2.0	11.2	53.1 ±0.7 [†]	7.8	58.8 ±0.5 [†]	5.4
			36.3 ±1.1 [†]	17.0	46.0 ±0.5 [†]	11.2	50.9 ±1.5	8.2	58.5 ±0.8	5.0

Table 4: IEMOCAP results. Metric: Weighted-Fscore averaged over 10 random runs. BE = Best Epoch. Results span across different amount of available training data. Validation and testing splits are fixed across configurations. [†] represents significant difference with $p < 0.05$ over randomly initialized model as per two-tailed Wilcoxon rank sum hypothesis test [80].

Variant	Initial Weights		Dataset: DailyDialog			
			10%		100%	
	sent _{enc}	cxt _{enc}	F-score	BE	F-score	BE
(1)	-	-	33.5 ±2.2	12.3	45.3 ±1.9	7.9
(2)	θ^{BERT}	-	37.5 ±1.8	2.6	47.4 ±1.2	2.4
(3)	θ^{BERT}	θ_{cxt}^{ubuntu} $\theta_{cxt}^{cornell}$	37.7 ±3.1	3.1	47.1 ±.76	2.4
			38.5 ±1.5 [†]	3.2	48.0 ±1.8 [†]	2.4

Table 5: DailyDialog results. Metric: Weighted-Fscore averaged over 5 random runs. BE = Best Epoch. [†] represents significant difference with $p < 0.05$ over random initialized model as per two-tailed Wilcoxon rank sum hypothesis test [80].

4.4. Training Criteria

Hyper-parameter search. For each target dataset-model combination, we perform grid-search to select the appropriate hyper-parameters. In the search procedure, we keep the model architecture constant but vary learning rate (1e-3, 1e-4, and 1e-5), optimizer (Adam, RMSprop [78]), batch size (2-40 videos/batch), and dropout ({0.0, 0.5}). BERT-parameters contain dropout of 0.1 as in Devlin et al. [12]). For a particular dataset-model pair, the final hyper-parameter configuration is chosen based on the best performance on the respective validation set. In the case of negligible difference between the combinations, we use the Adam optimizer [79] as the default variant with $\beta = [0.9, 0.999]$ and learning rate $1e-4$.

Inference. We train our models on each target dataset for multiple runs (10:IEMOCAP, 5:DailyDialog, 5:SEMAINE). In each run, the training proceeds with an *early stopping* criterion of patience 10. During this training loop, the parameters with the least validation loss are finally chosen for the testing-set inference and evaluation.

5. Results and Analyses

Table 4 and 5 provide the performance results of ERC on classification datasets IEMOCAP and DailyDialog, respectively. In both the tables, we observe clear

Variant	Initial Weights		Dataset: SEMAINE							
			DV		DA		DP		DE	
	sent _{enc}	cxt _{enc}	r	BE	r	BE	r	BE	r	BE
(1)	-	-	0.14	4	0.27	6.2	0.18	12.8	-0.03	287.4
(2)	θ^{BERT}	-	0.64	13.8	0.36	7.8	0.33	4.8	-0.03	23
(3)	θ^{BERT}	θ_{cxt}^{ubuntu}	0.66	10.2	0.41	6	0.34	3.8	-0.03	23
		$\theta_{cxt}^{cornell}$	0.65	10.2	0.42	8.8	0.35	3.4	-0.029	22.7

Table 6: SEMAINE results. Metric (r): Pearson correlation coefficients averaged over 5 random runs. DV = Valence, DA = Activation/Arousal, DP = Power, DE =Anticipation/Expectation.

and statistically significant improvements of the models that use pre-trained weights over the randomly initialized variant. We see further improvements when context-modeling parameters from the source task (θ_{cxt}^{source}) are transferred, indicating the benefit of using TL in this context-level hierarchy.

Similar trends are observed in the regression task based on the SEMAINE corpus (see Table 6). For *valence*, *arousal*, and *power* dimensions, the improvement is significant. For *expectation*, the performance is marginally better but at a much lesser BE, indicating faster generalization.

In the following sections, we take a closer look at various aspects of our approach that include checking robustness towards limited-data scenarios, generalization time, and questioning design choices. We also provide additional analyses that probe the existence of data-split bias, domain influence, and effect of fine-tuning strategies.

5.1. Target Data Size

Present approaches in ERC primarily adopt supervised learning strategies that demand a high amount of annotated data. However, the publicly available datasets in this field belong to the small-to-medium range in the spectrum of dataset sizes. For example, other applications of NLP have datasets of much larger sizes – over 130k instances in SQuAD for Question Answering [81], over 393k instances in MNLI for language inference [82], and so on. This constraint inhibits the true potential of systems trained on these datasets. As a result, approaches that provide higher performance in a limited training-data scenario tend to be highly desirable, particularly for ERC.

We design experiments to check the robustness of our models against such limited settings. To limit the amount of available training data, we create random subsets of the training dialogues while maintaining the original label-distribution. In both Table 4 and 5, we observe that the pre-trained models are significantly more robust against limited training resources compared to models trained from scratch.

Effect of bias in random splits. We investigate the possibility of bias in the random splits, which aid in supporting our hypothesis. To eliminate this possibility, we further check if the improvement in our TL-based approach — for the limited-data scenarios — are triggered by such data-split bias. In other

words, we pose the following question, *if another training split is sampled from the original dataset, would our model provide similar improvements?* We provide evidence that this is indeed true.

Table 7 presents the results where for 10% and 50% training-data setup, we sample 4 independent splits from the IEMOCAP dataset. As seen from the table, different splits provide different results, which is expected owing to the variances in the samples and their corresponding labels. However, the relative performance within each split follows similar trends of improvement for TL-based models. This observation nullifies the potential existence of bias in the reported results.

5.2. Target Task’s Training Time

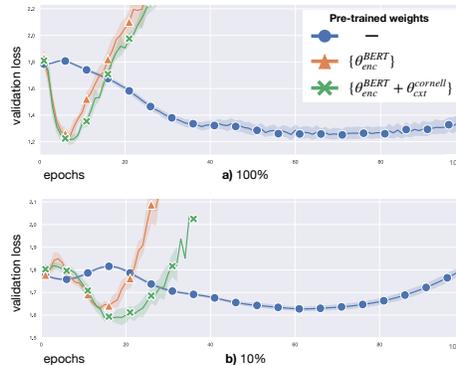


Figure 5: Validation loss across epochs in training for different weight-initialization settings on the IEMOCAP dataset. Part a) represents results when trained on 100% training data b) 10% training data split. For fair comparison, optimizer learning rates are fixed at $1e-4$.

In all the configurations in Table 4 and 5, we observe that the presence of weight initialization leads to faster convergence in terms of the best validation loss. Fig. 5 demonstrates the trace of the validation loss on training data configurations of the IEMOCAP dataset. As observed, the pre-trained models achieve their best epoch in a significantly shorter time which indicates that the transferred weights are helping the model better guide to its optimal performance.

5.3. Encoder Initialization

Table 8 provides a comparative study between the performance of models initialized with HRED-based sentence encoders (θ_{enc}^{source}) versus the BERT encoders (θ^{BERT}) that we use in our final networks. Results demonstrate that BERT provides better representations, which leads to better performance. Moreover, the positive effects of the context parameters are observed when coupled with the BERT encoders. This behavior indicates that the performance boosts provided by the context-encoders is contingent on the quality of sentence encoders. Observing this empirical evidence, we choose BERT-based sentence encoders in our final network.

Variant	Initial Weight		Dataset: IEMOCAP							
	sent _{enc}	cxt _{enc}	10%				50%			
			split ₁ *	split ₂	split ₃	split ₄	split ₁ *	split ₂	split ₃	split ₄
(1)	-	-	23.2 ±0.4	31.5 ±0.6	25.0 ±1.7	8.8 ±1.1	48.4 ±0.3	48.5 ±1.3	49.1 ±0.9	51.3 ±0.5
(2)	θ^{BERT}	-	32.4 ±1.1	31.6 ±1.2	30.5 ±0.8	23.65 ±1.3	49.2 ±1.0	49.0 ±0.7	48.8 ±0.9	51.4 ±0.6
(3)	θ^{BERT}	θ^{ubuntu}_{cxt}	35.7 ±1.1	32.0 ±1.1	39.0 ±0.2	24.90 ±3.0	53.1 ±0.7	53.2 ±1.3	52.9 ±1.9	54.2 ±0.8
		$\theta^{cornell}_{cxt}$	36.3 ±1.1	34.2 ±0.8	35.7 ±0.5	24.70 ±1.2	50.9 ±1.5	54.3 ±0.8	53.5 ±0.6	55.4 ±1.0

* primary split

Table 7: Table to investigate if split randomness incurs bias in results. Comparisons are held between two limited training data scenarios comprising 10% and 50% available training data. For both the cases, 4 independent splits are sampled and compared against. Metric: Weighted-Fscore averaged over 10 random runs.

Initial Weight		Dataset: IEMOCAP	
sent _{enc}	cxt _{enc}	10%	100%
		F-score	F-score
-	-	23.2 ±0.4	53.8 ±0.3
$\theta^{cornell}_{enc}$	-	26.3 ±0.9	54.9 ±0.3
	$\theta^{cornell}_{cxt}$	27.5 ±1.3	55.1 ±0.9
θ^{ubuntu}_{enc}	-	24.6 ±0.9	53.2 ±0.5
	θ^{ubuntu}_{cxt}	23.3 ±0.8	53.7 ±0.9
θ^{BERT}	-	32.4 ±1.1	55.1 ±0.6
	θ^{ubuntu}_{cxt}	35.7 ±1.1	58.8 ±0.5
	$\theta^{cornell}_{cxt}$	36.3 ±1.1	58.5 ±0.8

Table 8: Table to analyze HRED encoder vs BERT. Metric: Weighted-Fscore averaged over 10 random runs. BE = Best Epoch (average).

5.4. Impact of Source Domain.

We investigate if the choice of source datasets incur any significant change in the results. First, we define an emotional profile for the source datasets and observe whether any correlation is found between their emotive content versus the performance boost achieved by pre-training on them.

To set up an emotional profile, we look at the respective vocabularies of both corpora. For each token, we check its association with any emotion by using the emotion-lexicon provided by Mohammad and Turney [83]. The NRC Emotion Lexicon contains 6423 words belonging to emotion categories: *fear*, *trust*, *anger*, *sadness*, *anticipation*, *joy*, *surprise*, and *disgust*. It also assigns two broad categories: *positive* and *negative* to describe the type of connotation evoked by the words. We enumerate the frequency of each emotion category amongst the tokens of the source dataset’s vocabulary. To compose the vocabulary of both the source datasets, we set a minimum frequency threshold of 5, which provides 13518 and 18473 unique tokens for Cornell and Ubuntu, respectively. Each of the unique tokens is then lemmatized⁵ and cross-referenced with the lexicon, which provides 3099 (Cornell) and 2003 (Ubuntu) tokens with associated

⁵https://www.nltk.org/_modules/nltk/stem/wordnet.html

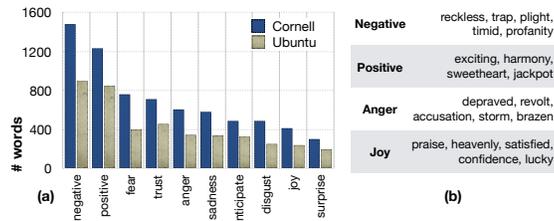


Figure 6: a) Frequency of emotive words from source datasets: Cornell and Ubuntu. b) randomly sampled words from Cornell associated to mentioned emotions.

Adapt Strategy	Iemocap	DD
Fixed weights	F-Score	F-Score
-	58.5	48.0
θ^{BERT}	17.0	32.1
$\theta^{BERT} + \theta_{cxt}^{cornell}$	9.3	4.5

Table 9: Average performance on ERC with pre-trained weights: $\{\theta^{BERT} + \theta_{cxt}^{cornell}\}$. Note: DD here means DailyDialog.

Models	Iemocap	SEMAINE			
		DV	DA	DP	DE
	F-Score	r	r	r	r
CNN	48.1	-0.01	0.01	-0.01	0.19
Memnet	55.1	0.16	0.24	0.23	0.05
c-LSTM	54.9	0.14	0.23	0.25	-0.04
c-LSTM + Att	56.1	0.16	0.25	0.24	0.10
CMN	56.1	0.23	0.29	0.26	-0.02
DialogueRNN	59.8	0.28	0.36	0.32	0.31
TL-ERC	58.8	0.66	0.42	0.35	-0.02

Table 10: Average performance of TL-ERC compared to previous state-of-the-art models.

emotions.

Fig. 6 presents the emotional profiles, which indicate that the Cornell dataset has a higher number of emotive tokens in its vocabulary. However, the results illustrated in Table 4, 5, and 6 *do not* present any significant difference between the two sources. A possible reason for this behavior attributes to the fact that such emotional profile relies on surface emotions derived from the vocabularies. However, as per our hypothesis, response generation includes emotional understanding as a latent process. This reasoning leads us to believe that surface emotions need not necessarily correlate to performance increments. Rather, the quality of generation would include such properties intrinsically.

5.5. Comparison with previous work.

Table 10 provides the results for various baselines detailed in Section 4.3. As seen, our proposed TL-ERC comfortably outperforms both non-contextual and contextual baselines. It achieves this without the aid of attention mechanisms that is used in c-LSTM + Att, multi-hop memory networks used in Memnet, and CMN. It also achieves competitive performance against DialogueRNN, which has three layers of inter-utterance recurrent layers, while TL-ERC has one. These trends indicate TL to be effective in our setup and provided promising directions for future research.

6. Challenges

In this section, we enlist the different challenges that we observed while experimenting with the proposed idea. These challenges provide roadmaps for further research on this topic to build better and robust systems.

6.1. Adaptation Strategies

We try two primary adaptation techniques used in inductive TL, *frozen* or *fine-tuned*. In the former setting, the borrowed weights are used for feature extraction, while in the latter, we train the weights along with the other new parameters of the target task’s model. Fine-tuning can also be performed using other techniques such as gradual unfreezing [84]. In Table 9, we experiment with freezing different amounts of transferred weights in our ERC model. We notice a degradation in performance with more frozen parameters. The datasets in ERC contain multi-class annotations with varying label distributions. With frozen parameters, our transferred model is unable to account for the label distribution and results in low recall for infrequent classes. We thus find the fine-tuning approach to be more effective in this setup.

However, fine-tuning all parameters also present higher susceptibility to over-fitting [20]. We observe this trait in Fig. 5, where the validation loss shoots up at a faster rate than the random counterpart. Finding a fine-balance in this trade-off remains an open problem.

6.2. Stability Issues

In the results, we observe that the variability of the models across multiple runs (in terms of the standard error) is relatively higher for the proposed models as compared to randomly initialized weights. Though, on average, our models perform significantly better, there remains a scope for improvement to achieve more stable training.

6.3. Variational Models

Many works utilize variational networks to model the uncertainties and variability in latent factors. For dialogue modeling, networks such as VHRED [85] incorporate such variational properties to model its latent processes. Emotional perception, in particular, has been argued to contain shades of multiple affective classes instead of a hard label assignment [86]. We, thus, posit that variational dialogue models such as VHRED also hold the potential for improving affective knowledge transfer.

We experiment on this concept by using VHRED as the source model. VHRED uses additional parameters to model its prior latent state \mathbf{z}_t , which is then concatenated with \mathbf{h}_t^{cxt} as follows:

$$\begin{aligned}\mathbf{h}_t^{enc} &= f_{\theta}^{enc}(\mathbf{x}_t) \\ \mathbf{h}_t^{cxt} &= f_{\theta}^{cxt}(\mathbf{h}_t^{enc}, \mathbf{h}_{t-1}^{cxt})\end{aligned}$$

Initial Weights	IEMOCAP F-Score	Dailydialog F-Score
HRED	58.5	48.0
VHRED	58.6	48.4

Table 11: Average performance on ERC with pre-trained weights: $\{\theta^{BERT} + \theta_{cxt}^{cornell}\}$ for VHRED, $\theta_{cxt}^{cornell}$ contain additional parameters modeling the latent prior state.

Generative Training	IEMOCAP F-Score	Dailydialog F-Score
Source	58.5	48.0
Source + Target	58.0	47.2

Table 12: Average performance on ERC with pre-trained weights: $\{\theta^{BERT} + \theta_{cxt}^{cornell}\}$.

$$\begin{aligned}
p_{\theta}(\mathbf{z}_t | \mathbf{x}_{\leq t}) &= \mathcal{N}(\mathbf{z}_t | \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t \mathbf{I}) \\
\text{where } \boldsymbol{\mu}_t &= \text{MLP}_{\theta}(\mathbf{h}_t^{cxt}) \\
\boldsymbol{\sigma}_t &= \text{Softplus}(\text{MLP}_{\theta}(\mathbf{h}_t^{cxt})) \\
\mathbf{h}_t^{cxt} &= [\mathbf{h}_t^{cxt}, \mathbf{z}_t]
\end{aligned}$$

As a result, our set of transferred parameters contain the additional parameters of MLP_{θ} , included in θ_{cxt}^{source} . Table 11 presents the result of using VHRED parameters. Unfortunately, we do not find significant difference between the parameters from VHRED as opposed to HRED. However, the lack of degradation in the results promise possible future improvements in such designs.

6.4. In-domain Generative Fine-Tuning

We try in-domain tuning of the generative HRED model by performing conversation modeling on the ERC resources. Finally, we transfer these re-tuned weights for the discriminative ERC task. However, we do not find this procedure to be helpful (Table 12). TL between generative tasks, especially with small-scale target resources, is a challenging task. As a result, we find sub-optimal generation in ERC datasets whose further transfer for the classification does not provide any improvement.

6.5. Quality of Generative Models

Despite their recent developments, generative dialogue models still suffer from numerous shortcomings. Challenges include lack of diversity in the responses, which results in the generation of universal sentences, such as *I don't know* [87, 88]. Coherence in topic and emotions are also difficult to maintain while generating responses [89]. Similar traits are observed in our pre-training experiments.

Although TL-ERC obtains significant improvement in the results, we obtain it with a simple dialogue model. We, thus, believe that further improvements are possible and is contingent on the quality of the dialogue generator. As research in dialogue systems inch towards the *perfect dialogue generator*, it would also benefit ERC via our proposed TL-ERC framework.

7. Conclusion

In this paper, we presented a novel framework of transfer learning (TL-ERC) for ERC that uses pre-trained affective information from dialogue generators. We presented various experiments with different scenarios to investigate the effect of this procedure. We found that using such pre-trained weights help the overall task and also provide added benefits in terms of lesser training epochs for good generalization. We primarily experimented on dyadic conversations both in the source and the target tasks. In the future, we aim to investigate the more general setting of multi-party conversations. This setting will increase the complexity of the task, as pre-training would require multi-party data and special training schemes to capture complex influence dynamics.

Code used for this work is publicly available at <https://github.com/SenticNet/conv-emotion>.

Acknowledgement

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE’s official grant number T1 251RES1820. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research.

References

References

- [1] S. Poria, N. Majumder, R. Mihalcea, E. H. Hovy, Emotion recognition in conversation: Research challenges, datasets, and recent advances, *IEEE Access* 7 (2019) 100943–100953. URL: <https://doi.org/10.1109/ACCESS.2019.2929050>. doi:10.1109/ACCESS.2019.2929050.
- [2] H. Chen, X. Liu, D. Yin, J. Tang, A survey on dialogue systems: Recent advances and new frontiers, *SIGKDD Explorations* 19 (2017) 25–35. URL: <https://doi.org/10.1145/3166054.3166058>. doi:10.1145/3166054.3166058.
- [3] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 2018, pp. 2122–2132. URL: <https://www.aclweb.org/anthology/N18-1193/>.
- [4] S. J. Pan, Q. Yang, A survey on transfer learning, *IEEE Trans. Knowl. Data Eng.* 22 (2010) 1345–1359. URL: <https://doi.org/10.1109/TKDE.2009.191>. doi:10.1109/TKDE.2009.191.

- [5] E. Hovy, Generating natural language under pragmatic constraints, *Journal of Pragmatics* 11 (1987) 689–719.
- [6] E. Weigand, Emotions in dialogue, *Dialoganalyse VI/1: Referate der 6. Arbeitstagung*, Prag 1996 16 (2017) 35.
- [7] J. Sidnell, T. Stivers, *The handbook of conversation analysis*, volume 121, John Wiley & Sons, 2012.
- [8] P. Koval, P. Kuppens, Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia., *Emotion* 12 (2012) 256.
- [9] C. Navarretta, Mirroring facial expressions and emotions in dyadic conversations, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, Portorož, Slovenia, May 23-28, 2016., 2016. URL: <http://www.lrec-conf.org/proceedings/lrec2016/summaries/258.html>.
- [10] T. Shimizu, N. Shimizu, H. Kobayashi, Pretraining sentiment classifiers with unlabeled dialog data, in: I. Gurevych, Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers, Association for Computational Linguistics, 2018, pp. 764–770. URL: <https://www.aclweb.org/anthology/P18-2121/>. doi:10.18653/v1/P18-2121.
- [11] C. Danescu-Niculescu-Mizil, L. Lee, Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs, in: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, Association for Computational Linguistics, 2011, pp. 76–87.
- [12] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: [90], 2019, pp. 4171–4186. URL: <https://www.aclweb.org/anthology/N19-1423/>.
- [13] R. K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, *J. Mach. Learn. Res.* 6 (2005) 1817–1853. URL: <http://jmlr.org/papers/v6/ando05a.html>.
- [14] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2, 2019, Tutorial Abstracts, 2019, pp. 15–18. URL: <https://www.aclweb.org/anthology/N19-5004/>.
- [15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: *Advances*

- in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., 2013, pp. 3111–3119.
- [16] B. McCann, J. Bradbury, C. Xiong, R. Socher, Learned in translation: Contextualized word vectors, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 6294–6305. URL: <http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors>.
- [17] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 2227–2237. URL: <https://www.aclweb.org/anthology/N18-1202/>.
- [18] A. M. Dai, Q. V. Le, Semi-supervised sequence learning, in: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 3079–3087. URL: <http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning>.
- [19] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, CoRR abs/1906.08237 (2019). URL: <http://arxiv.org/abs/1906.08237>. arXiv:1906.08237.
- [20] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 328–339. URL: <https://www.aclweb.org/anthology/P18-1031/>. doi:10.18653/v1/P18-1031.
- [21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pre-training approach, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [22] L. Chen, A. Moschitti, Transfer learning for sequence labeling using source model and target data, arXiv preprint arXiv:1902.05309 (2019).

- [23] M. Qiu, L. Yang, F. Ji, W. Zhou, J. Huang, H. Chen, B. Croft, W. Lin, Transfer learning for context-aware question matching in information-seeking conversations in e-commerce, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 208–213.
- [24] C. Wu, S. C. H. Hoi, R. Socher, C. Xiong, Tod-bert: Pre-trained natural language understanding for task-oriented dialogues, CoRR abs/2004.06871 (2020). URL: <https://arxiv.org/abs/2004.06871>. arXiv:2004.06871.
- [25] M. Henderson, I. Casanueva, N. Mrksic, P. Su, T. Wen, I. Vulic, Convert: Efficient and accurate conversational representations from transformers, CoRR abs/1911.03688 (2019). URL: <http://arxiv.org/abs/1911.03688>. arXiv:1911.03688.
- [26] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: <https://openreview.net/forum?id=H1eA7AetvS>.
- [27] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: <https://doi.org/10.18653/v1/D19-1410>. doi:10.18653/v1/D19-1410.
- [28] Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, Dialogpt: Large-scale generative pre-training for conversational response generation, CoRR abs/1911.00536 (2019). URL: <http://arxiv.org/abs/1911.00536>. arXiv:1911.00536.
- [29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (2019) 9.
- [30] E. Cambria, S. Poria, A. Hussain, B. Liu, Computational intelligence for affective computing and sentiment analysis [guest editorial], IEEE Comput. Intell. Mag. 14 (2019) 16–17. URL: <https://doi.org/10.1109/MCI.2019.2901082>. doi:10.1109/MCI.2019.2901082.
- [31] M. Imani, G. A. Montazer, A survey of emotion recognition methods with emphasis on e-learning environments, J. Netw. Comput. Appl. 147 (2019). URL: <https://doi.org/10.1016/j.jnca.2019.102423>. doi:10.1016/j.jnca.2019.102423.
- [32] J. Liscombe, G. Riccardi, D. Hakkani-Tür, Using context to improve emotion detection in spoken dialog systems, in: INTERSPEECH 2005 - Eurospeech,

- 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005, ISCA, 2005, pp. 1845–1848. URL: http://www.isca-speech.org/archive/interspeech_2005/i05_1845.html.
- [33] S. N. Schiaffino, A. Amandi, Intelligent user profiling, in: M. Bramer (Ed.), *Artificial Intelligence: An International Perspective*, volume 5640 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 193–216. URL: https://doi.org/10.1007/978-3-642-03226-4_11. doi:10.1007/978-3-642-03226-4_11.
- [34] S. Li, W. Deng, Deep facial expression recognition: A survey, CoRR abs/1804.08348 (2018). URL: <http://arxiv.org/abs/1804.08348>. arXiv:1804.08348.
- [35] S. Wang, P. Phillips, Z. Dong, Y. Zhang, Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm, *Neurocomputing* 272 (2018) 668–676. URL: <https://doi.org/10.1016/j.neucom.2017.08.015>. doi:10.1016/j.neucom.2017.08.015.
- [36] G. Drakopoulos, G. Pikramenos, E. D. Spyrou, S. J. Perantonis, Emotion recognition from speech: A survey, in: A. Bozzon, F. J. D. Mayo, J. Filipe (Eds.), *Proceedings of the 15th International Conference on Web Information Systems and Technologies, WEBIST 2019, Vienna, Austria, September 18-20, 2019*, ScitePress, 2019, pp. 432–439.
- [37] C. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, *Artif. Intell. Rev.* 43 (2015) 155–177. URL: <https://doi.org/10.1007/s10462-012-9368-5>. doi:10.1007/s10462-012-9368-5.
- [38] C. Maréchal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, K. Wegrzyn-Wolska, Survey on ai-based multimodal methods for emotion detection, in: J. Kolodziej, H. González-Vélez (Eds.), *High-Performance Modelling and Simulation for Big Data Applications - Selected Results of the COST Action IC1406 cHiPSet*, volume 11400 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 307–324. URL: https://doi.org/10.1007/978-3-030-16272-6_11. doi:10.1007/978-3-030-16272-6_11.
- [39] C. O. Alm, D. Roth, R. Sproat, Emotions from text: Machine learning for text-based emotion prediction, in: *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada, The Association for Computational Linguistics, 2005*, pp. 579–586. URL: <https://www.aclweb.org/anthology/H05-1073/>.
- [40] C. Strapparava, R. Mihalcea, Learning to identify emotions in text, in: R. L. Wainwright, H. Haddad (Eds.), *Proceedings of the 2008 ACM Symposium*

- on Applied Computing (SAC), Fortaleza, Ceara, Brazil, March 16-20, 2008, ACM, 2008, pp. 1556–1560. URL: <https://doi.org/10.1145/1363686.1364052>. doi:10.1145/1363686.1364052.
- [41] C. Strapparava, A. Valitutti, Wordnet affect: an affective extension of wordnet, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal, European Language Resources Association, 2004. URL: <http://www.lrec-conf.org/proceedings/lrec2004/summaries/369.htm>.
- [42] A. Esuli, F. Sebastiani, SENTIWORDNET: A publicly available lexical resource for opinion mining, in: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, D. Tapias (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006, European Language Resources Association (ELRA), 2006, pp. 417–422. URL: <http://www.lrec-conf.org/proceedings/lrec2006/summaries/384.html>.
- [43] W. Wang, L. Chen, K. Thirunarayan, A. P. Sheth, Harnessing twitter "big data" for automatic emotion identification, in: 2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Conference on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012, IEEE Computer Society, 2012, pp. 587–592. URL: <https://doi.org/10.1109/SocialCom-PASSAT.2012.119>. doi:10.1109/SocialCom-PASSAT.2012.119.
- [44] W. Y. Choi, K. Y. Song, C. W. Lee, Convolutional attention networks for multimodal emotion recognition from speech and text data, in: Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 28–34. URL: <https://www.aclweb.org/anthology/W18-3304>. doi:10.18653/v1/W18-3304.
- [45] V. Chernykh, G. Sterling, P. Prihodko, Emotion recognition from speech with recurrent neural networks, CoRR abs/1701.08071 (2017). URL: <http://arxiv.org/abs/1701.08071>. arXiv:1701.08071.
- [46] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017, IEEE, 2017, pp. 2227–2231. URL: <https://doi.org/10.1109/ICASSP.2017.7952552>. doi:10.1109/ICASSP.2017.7952552.
- [47] M. Dragoni, An evolutionary strategy for concept-based multi-domain sentiment analysis, IEEE Comput. Intell. Mag. 14 (2019) 18–27. URL: <https://doi.org/10.1109/MCI.2019.2901083>. doi:10.1109/MCI.2019.2901083.

- [48] J. Han, Z. Zhang, B. W. Schuller, Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives [review article], *IEEE Comput. Intell. Mag.* 14 (2019) 68–81. URL: <https://doi.org/10.1109/MCI.2019.2901088>. doi:10.1109/MCI.2019.2901088.
- [49] J. Yu, L. Marujo, J. Jiang, P. Karuturi, W. Brendel, Improving multi-label emotion classification via sentiment classification with dual attention transfer network, in: [91], 2018, pp. 1097–1102. URL: <https://www.aclweb.org/anthology/D18-1137/>.
- [50] G. Daval-Frerot, A. Boucekif, A. Moreau, Epita at semeval-2018 task 1: Sentiment analysis using transfer learning approach, in: *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, 2018, pp. 151–155. URL: <https://www.aclweb.org/anthology/S18-1021/>.
- [51] A. Boucekif, P. Joshi, L. Boucekif, H. Afli, Epita-adapt at semeval-2019 task 3: Detecting emotions in textual conversations using deep learning models combination, in: *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 215–219.
- [52] H. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Z. Zhang, P. Cohen, D. Bohus, R. Horaud, H. Meng (Eds.), *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, Seattle, WA, USA, November 09 - 13, 2015, ACM, 2015, pp. 443–449. URL: <http://doi.acm.org/10.1145/2818346.2830593>. doi:10.1145/2818346.2830593.
- [53] J. Deng, Z. Zhang, E. Marchi, B. W. Schuller, Sparse autoencoder-based feature transfer learning for speech emotion recognition, in: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013, Geneva, Switzerland, September 2-5, 2013*, IEEE Computer Society, 2013, pp. 511–516. URL: <https://doi.org/10.1109/ACII.2013.90>. doi:10.1109/ACII.2013.90.
- [54] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, S. Lehmann, Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in: M. Palmer, R. Hwa, S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Association for Computational Linguistics, 2017, pp. 1615–1625. URL: <https://doi.org/10.18653/v1/d17-1169>. doi:10.18653/v1/d17-1169.
- [55] A. V. González-Garduño, V. P. B. Hansen, J. Bingel, I. Augenstein, A. Søgaard, Coastal at semeval-2019 task 3: Affect classification in dialogue using attentive bilstms, in: *Proceedings of the 13th International*

- Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, 2019, pp. 169–174. URL: <https://www.aclweb.org/anthology/S19-2026/>.
- [56] W. Jiao, H. Yang, I. King, M. R. Lyu, Higr: Hierarchical gated recurrent units for utterance-level emotion recognition, in: [90], 2019, pp. 397–406. URL: <https://www.aclweb.org/anthology/N19-1037/>.
- [57] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: interactive conversational memory network for multimodal emotion detection, in: [91], 2018, pp. 2594–2604. URL: <https://www.aclweb.org/anthology/D18-1280/>.
- [58] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. F. Gelbukh, Dialoguecn: A graph convolutional neural network for emotion recognition in conversation, CoRR abs/1908.11540 (2019). URL: <http://arxiv.org/abs/1908.11540>. arXiv:1908.11540.
- [59] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, G. Zhou, Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations, in: [92], 2019, pp. 5415–5421. URL: <https://doi.org/10.24963/ijcai.2019/752>. doi:10.24963/ijcai.2019/752.
- [60] Y. Zhang, Q. Li, D. Song, P. Zhang, P. Wang, Quantum-inspired interactive networks for conversational sentiment analysis, in: [92], 2019, pp. 5436–5442. URL: <https://doi.org/10.24963/ijcai.2019/755>. doi:10.24963/ijcai.2019/755.
- [61] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. F. Gelbukh, E. Cambria, Dialoguernn: An attentive RNN for emotion detection in conversations, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019., AAAI Press, 2019, pp. 6818–6825. URL: <https://aaai.org/ojs/index.php/AAAI/article/view/4657>.
- [62] P. Zhong, D. Wang, C. Miao, Knowledge-enriched transformer for emotion detection in textual conversations, CoRR abs/1909.10681 (2019). URL: <http://arxiv.org/abs/1909.10681>. arXiv:1909.10681.
- [63] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, Semeval-2019 task 3: Emocontext contextual emotion detection in text, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, Association for Computational Linguistics, 2019, pp. 39–48. URL: <https://doi.org/10.18653/v1/s19-2005>. doi:10.18653/v1/s19-2005.

- [64] Y. Huang, S. Lee, M. Ma, Y. Chen, Y. Yu, Y. Chen, Emotionx-idea: Emotion BERT - an affectional model for conversation, CoRR abs/1908.06264 (2019). URL: <http://arxiv.org/abs/1908.06264>. arXiv:1908.06264.
- [65] W. Jiao, M. R. Lyu, I. King, Pt-code: Pre-trained context-dependent encoder for utterance-level emotion recognition, CoRR abs/1910.08916 (2019). URL: <http://arxiv.org/abs/1910.08916>. arXiv:1910.08916.
- [66] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, in: D. Schuurmans, M. P. Wellman (Eds.), Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., AAAI Press, 2016, pp. 3776–3784. URL: <http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957>.
- [67] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.
- [68] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, 2017, pp. 873–883. URL: <https://doi.org/10.18653/v1/P17-1081>. doi:10.18653/v1/P17-1081.
- [69] R. Lowe, N. Pow, I. Serban, J. Pineau, The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, in: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2015, pp. 285–294.
- [70] Y. Park, J. Cho, G. Kim, A hierarchical latent structure for variational conversation modeling, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1792–1801.
- [71] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335.
- [72] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, S. Niu, Dailydialog: A manually labelled multi-turn dialogue dataset, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 986–995.
- [73] B. W. Schuller, M. F. Valstar, F. Eyben, R. Cowie, M. Pantic, AVEC 2012: the continuous audio/visual emotion challenge, in: L. Morency,

- D. Bohus, H. K. Aghajan, J. Cassell, A. Nijholt, J. Epps (Eds.), International Conference on Multimodal Interaction, ICMI '12, Santa Monica, CA, USA, October 22-26, 2012, ACM, 2012, pp. 449–456. URL: <https://doi.org/10.1145/2388676.2388776>. doi:10.1145/2388676.2388776.
- [74] Y. Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1746–1751. URL: <https://doi.org/10.3115/v1/d14-1181>. doi:10.3115/v1/d14-1181.
- [75] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 2440–2448. URL: <http://papers.nips.cc/paper/5846-end-to-end-memory-networks>.
- [76] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (1997) 1735–1780. URL: <https://doi.org/10.1162/neco.1997.9.8.1735>. doi:10.1162/neco.1997.9.8.1735.
- [77] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: V. Raghavan, S. Aluru, G. Karypis, L. Miele, X. Wu (Eds.), 2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017, IEEE Computer Society, 2017, pp. 1033–1038. URL: <https://doi.org/10.1109/ICDM.2017.134>. doi:10.1109/ICDM.2017.134.
- [78] T. Tieleman, G. Hinton, Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude, COURSE: Neural Networks for Machine Learning, 2012.
- [79] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: <http://arxiv.org/abs/1412.6980>.
- [80] N. Nachar, et al., The mann-whitney u: A test for assessing whether two independent samples come from the same distribution, *Tutorials in quantitative Methods for Psychology* 4 (2008) 13–20.
- [81] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for squad, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers,

- Association for Computational Linguistics, 2018, pp. 784–789. URL: <https://www.aclweb.org/anthology/P18-2124/>. doi:10.18653/v1/P18-2124.
- [82] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 1112–1122. URL: <http://aclweb.org/anthology/N18-1101>.
- [83] S. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, *Computational Intelligence* 29 (2013) 436–465. URL: <https://doi.org/10.1111/j.1467-8640.2012.00460.x>. doi:10.1111/j.1467-8640.2012.00460.x.
- [84] M. E. Peters, S. Ruder, N. A. Smith, To tune or not to tune? adapting pretrained representations to diverse tasks, in: I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, M. Rei (Eds.), Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019., Association for Computational Linguistics, 2019, pp. 7–14. URL: <https://www.aclweb.org/anthology/W19-4302/>.
- [85] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, Y. Bengio, A hierarchical latent variable encoder-decoder model for generating dialogues, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., 2017, pp. 3295–3301. URL: <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567>.
- [86] E. Mower, M. J. Mataric, S. S. Narayanan, A framework for automatic human emotion classification using emotion profiles, *IEEE Trans. Audio, Speech & Language Processing* 19 (2011) 1057–1070. URL: <https://doi.org/10.1109/TASL.2010.2076804>. doi:10.1109/TASL.2010.2076804.
- [87] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 110–119. URL: <https://www.aclweb.org/anthology/N16-1014>. doi:10.18653/v1/N16-1014.
- [88] Y. Song, C. Li, J. Nie, M. Zhang, D. Zhao, R. Yan, An ensemble of retrieval-based and generation-based human-computer conversation systems, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden., 2018, pp. 4382–4388. URL: <https://doi.org/10.24963/ijcai.2018/609>. doi:10.24963/ijcai.2018/609.

- [89] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 2018, pp. 730–739. URL: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16455>.
- [90] J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019. URL: <https://www.aclweb.org/anthology/volumes/N19-1/>.
- [91] E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018. URL: <https://www.aclweb.org/anthology/volumes/D18-1/>.
- [92] S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019. URL: <https://doi.org/10.24963/ijcai.2019>. doi:10.24963/ijcai.2019.