# CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts

**Peixiang Zhong,**[1,2] **Di Wang,** [2] **Pengfei Li,** [3] **Chen Zhang,** [4] **Hao Wang,** [4*] **Chunyan Miao** [1,2*]

[1] Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University (NTU), Singapore
[2] Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, NTU, Singapore
[3] School of Electrical and Electronic Engineering, NTU, Singapore
[4] Alibaba Group, China
{peixiang001, wangdi, pli006, ascymiao}@ntu.edu.sg, zhangchen010295@163.com, cashenry@126.com

## Abstract

Rationality and emotion are two fundamental elements of humans. Endowing agents with rationality and emotion has been one of the major milestones in AI. However, in the field of conversational AI, most existing models only specialize in one aspect and neglect the other, which often leads to dull or unrelated responses. In this paper, we hypothesize that combining rationality and emotion into conversational agents can improve response quality. To test the hypothesis, we focus on one fundamental aspect of rationality, i.e., commonsense, and propose CARE, a novel model for commonsense-aware emotional response generation. Specifically, we first propose a framework to learn and construct commonsense-aware emotional latent concepts of the response given an input message and a desired emotion. We then propose three methods to collaboratively incorporate the latent concepts into response generation. Experimental results on two large-scale datasets support our hypothesis and show that our model can produce more accurate and commonsense-aware emotional responses and achieve better human ratings than state-of-the-art models that only specialize in one aspect.

## Introduction

Rationality and emotion are two fundamental elements of humans and indispensable to our social interactions (Keltner and Haidt 1999; Colman 2003). Endowing agents with rationality and emotion has been one of the major milestones in AI. In recent studies of conversational AI, there is an emerging research trend in endowing conversational models with rationality (Zhou et al. 2018b; Zhang et al. 2020) or emotion (Zhou et al. 2018a; Song et al. 2019). Rational conversational models can leverage commonsense knowledge to reason and have been shown to produce more appropriate and informative responses (Zhou et al. 2018b). Emotional conversational models can generate appropriate emotional responses, leading to improved user satisfaction (Prendinger and Ishizuka 2005; Zhou et al. 2018a) and long-term relationships with users (Zhou et al. 2018c).

Rationality and emotion are not independent for humans (De Sousa 1990). In fact, emotions are often rational in

| **Model** | **Message** | Why do you avoid starbucks? |
|---|---|---|
| | **Gold** | I don't like their coffee. |
| EmoDS | Neutral | I'm not a gamer so it's just that I'm a bit more tired of them. |
| | Joy | Yeah, that's pretty cool. I like that. |
| | Sadness | That sucks. |
| Ours (CARE) | Neutral | It's just a general dislike of the **company**. |
| | Joy | **Tea** is my *favorite*. |
| | Sadness | I'm not a big fan of **coffee**. It 's too *bitter* for me. |

Table 1: Sample responses from EmoDS (Song et al. 2019) and our model. EmoDS generates generic or unrelated emotional responses. Our model extracts the message concept "starbucks", and generates more commonsense-aware emotional responses by referring to our constructed relational (in bold) and emotional (in italic) latent concepts, e.g., company, coffee and bitter.

social interactions (Pham 2007). However, most existing conversational models only specialize in one aspect and neglect the other[1], which often leads to dull or unrelated responses. For example, as shown in Table 1, the state-of-the-art emotional conversational model (EmoDS) (Song et al. 2019) produces generic or unrelated emotional responses due to the lack of specific modelling of rationality. In addition, existing rational conversational models, e.g., the commonsense-aware conversational model (CCM) (Zhou et al. 2018b), are not able to generate emotional responses, rendering them difficult to build long-term relationships with users (Zhou et al. 2018c).

A recent work (Roller et al. 2020) proposed to blend several human skills such as knowledge, personality, and empathy into a conversational agent and obtained the state-of-the-art performance in human evaluations. Their experimental analysis suggests that blending these skills is critical for achieving good human ratings. Motivated by the facts that 1) rationality and emotion are two fundamental qualities of humans and that 2) empirical performance improvement has been achieved via combining several human qualities

*Corresponding author


[1]One exception is XiaoIce (Zhou et al. 2018c); however, it has no public API and only supports Mandarin.

(Roller et al. 2020), we hypothesize that combining rationality and emotion into conversational agents can improve response quality and their human ratings.

In this paper, we narrow the scope of rationality and emotion to specific settings for easier implementation and evaluation. Specifically, we focus on one fundamental aspect of rationality, i.e., commonsense, and the discrete representation of emotion. Commonsense is an important foundation of rationality and the basis of rational human conversations (Ross 1978). The discrete representation of emotion categorizes emotions into discrete basic emotions, e.g., joy, anger, etc., and is a well-established emotion theory in Psychology (Ekman 1992). To test our hypothesis, we propose a novel model for **C**ommonsense-**A**ware **R**esponse generation with specified **E**motions (**CARE**) and assess its empirical performance. Two major challenges to this task are 1) the lack of relevant datasets or resources that can provide such supervision and 2) how to generate appropriate commonsense-aware emotional words. We tackle the first challenge by building an emotion-aware commonsense knowledge graph (**EA-CKG**) to integrate commonsense and emotion knowledge. We tackle the second challenge by incorporating both relational and emotional latent concepts constructed from EA-CKG into response generation. Specifically, we build EA-CKG by augmenting an external CKG with emotional triplets extracted from emotional conversations. We then construct latent concepts using learned EA-CKG embeddings, endowing the response with commonsense and emotion by reasoning over the EA-CKG. Finally, we propose three methods to sequentially and collaboratively incorporate the latent concepts during attention, optimization, and sampling. CARE is illustrated in Figure 1.

In summary, our contributions are as follows:

- We identify the problem of lacking either rationality or emotion in existing conversational models, which often leads to dull or unrelated responses. We hypothesize that combining rationality and emotion into conversational agents can improve response quality.
- We focus on one fundamental aspect of rationality, i.e., commonsense, and propose CARE, the first commonsense-aware emotional response generation model, to address the aforementioned problem.
- We conduct extensive automatic and human evaluations and show that CARE can produce better commonsense-aware emotional responses than state-of-the-art models that only specialize in one aspect. The experimental results support our hypothesis.

## Related Work

**Rational Response Generation**: Existing rational response generation models usually rely on knowledge bases, such as open-domain response generation (Han et al. 2015; Young et al. 2018; Ghazvininejad et al. 2018; Liu et al. 2018; Tuan, Chen, and Lee 2019; Moon et al. 2019), task-oriented response generation (Madotto, Wu, and Fung 2018; Wu, Socher, and Xiong 2019) and question answering (Sun et al. 2018; Banerjee et al. 2019). Zhou et al. (2018b) proposed CCM to incorporate commonsense knowledge by applying attention mechanisms on 1-hop knowledge triplets for open-domain response generation. Zhang et al. (2020) proposed ConceptFlow to extend CCM to multi-hop knowledge triplets. Different from CCM and ConceptFlow, our model is not restricted by the coverage of the CKG and can learn novel knowledge triplets for response generation.

**Emotional Response Generation**: Emotional conversational models (Hasegawa et al. 2013; Asghar et al. 2018; Zhou and Wang 2018; Zhong, Wang, and Miao 2019a; Rashkin et al. 2019; Lin et al. 2019) are also emerging. Zhou et al. (2018a) extended the Seq2Seq model by proposing an internal memory module to capture emotional state changes and an external memory module to generate emotional words. Song et al. (2019) proposed an emotion classifier to guide the response generation. In contrast, our model generates emotional responses by leveraging emotional latent concepts constructed from KG embeddings.

**Controlled Text Generation**: Recent controlled text generation methods are primarily based on generative adversarial networks (GAN) (Hu et al. 2017; Li and Tuzhilin 2019), language models (Ghosh et al. 2017) and Seq2Seq models (Xing et al. 2017; Xu et al. 2019). Keskar et al. (2019) trained a Transformer-based conditional language model on a large collection of corpora with control codes that govern style, content, and task-specific behavior. Li and Sun (2018) and Peng et al. (2019) proposed topic-aware emotional response generation models. In contrast, we focus on commonsense, i.e., the semantic network of words, instead of topics, i.e., word clusters.

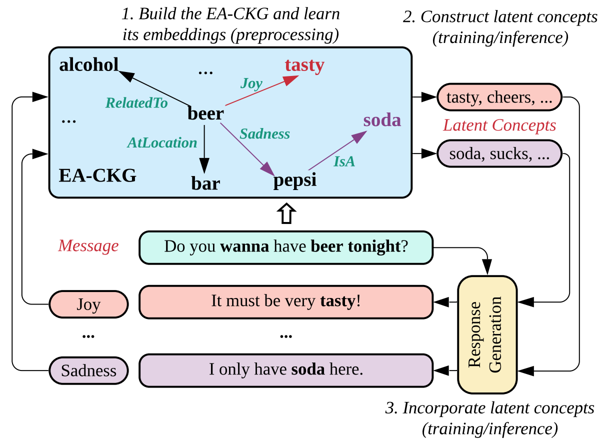


Figure 1: Illustration of CARE. Given the message "Do you wanna have beer tonight?" ("beer" is a message concept) and the learned EA-CKG embeddings, CARE first constructs latent concepts depending on the specified emotions of the response. For example, "tasty" is constructed for "joy" and "soda" is constructed for "sadness", because "tasty" is linked to "beer" via the "joy" relation, and "soda" is linked to "beer" via a composite of "sadness" and "IsA" relations. Then CARE leverages the proposed three methods to incorporate the latent concepts, e.g., "tasty", into response generation.

## Our CARE Model

In this section, we introduce the task definition and our CARE model, which includes a framework for constructing latent concepts and three methods to incorporate the latent concepts.

### Task Definition

We denote $\{X_i, Y_i, e_i\}, i = 1, ..., N$, as a collection of $\{$*message*, *response*, *emotion*$\}$ tuples, where $e_i$ is chosen from a predefined set of emotions and denotes the emotion category of $Y_i$, and $N$ denotes the number of conversations in the training dataset. Our task can be formulated as follows: given a new message $X_{\text{new}}$ and an emotion category $e$, generate a natural and commonsense-aware response $Y_{\text{new}}$ that has emotion $e$.

### Latent Concepts Construction Framework

In this framework, we first build an emotion-aware commonsense knowledge graph (EA-CKG) and then construct latent concepts from EA-CKG.

**EA-CKG** We extract emotional triplets from emotional conversations and augment them into an external CKG to obtain EA-CKG. We use ConceptNet (Speer, Chin, and Havasi 2017) as our CKG[2]. Each triplet in ConceptNet follows the $\{$*head*, *relation*, *tail*$\}$ format, e.g., $\{$*beer*, *AtLocation*, *bar*$\}$. Note that we use n-gram matching with ConceptNet to extract concepts from utterances, and ignore stopwords and n-grams that are formed entirely by stopwords. We define an emotional triplet as in the $\{$*msg_concept*, *emotion*, *res_concept*$\}$ format, representing an emotional link from a message concept to a response concept. For example, given a message "I heard there is a bar nearby with nice beer." and its response "I love tasty beer." with joy emotion, the triplet $\{$*beer*, *joy*, *tasty*$\}$ is a valid emotional triplet because there is a commonly expressed emotional link, i.e., *joy*, from *beer* in the message to *tasty* in the response.

We propose a two-step approach based on the pointwise mutual information (PMI) (Church and Hanks 1990) to extract such emotional triplets from emotional conversations. PMI can measure the association between two words in a corpus. We extend the smoothed positive PMI, i.e., $\text{PPMI}_\alpha$ (Levy, Goldberg, and Dagan 2015), as follows:

$$\text{PPMI}_\alpha(w_1, w_2) = \max\left(\log_2 \frac{P(w_1, w_2)}{P_\alpha(w_1)P_\alpha(w_2)}, 0\right), \quad (1)$$

where $(w_1, w_2)$ denotes the word pair, $P_\alpha(w) = \frac{\text{count}(w)^\alpha}{\sum_x \text{count}(x)^\alpha}$ denotes the smoothed probability of $w$, and $\alpha$ denotes a smoothing factor set to $0.75$ (Levy, Goldberg, and Dagan 2015) to alleviate the bias towards rare words.

In our two-step approach, we first construct a PPMI matrix between concepts in messages and in the corresponding responses to extract strongly associated concept pairs in conversations[3], denoted as conversational concept pairs (**CCP**). Note that in this case, $w_1$ refers to a message concept and $w_2$ refers to a response concept in Equation 1. We then construct a second PPMI matrix between CCP and their expressed emotions and extract CCP that statistically express certain emotions more often than other emotions[4]. Note that in this case, $w_1$ refers to a CCP and $w_2$ refers to its expressed emotion in Equation 1. We do not smooth $P(w_2)$. By using this two-step approach, we can effectively extract conversational triplets that are commonly expressed with certain emotions. The statistics of EA-CKG are presented in Table 2. Our approach shares similarities with commonsense knowledge base completion methods (Li et al. 2016c; Saito et al. 2018; Bosselut et al. 2019); however, they cannot be trivially adapted to extract emotional CCP.

| CKG | #entity | #relation | #triplet |
|---|---|---|---|
| ConceptNet | 182K | 36 | 1.48M |
| EA-CKG (Reddit) | 182K | 42 | 1.58M |
| EA-CKG (Twitter) | 182K | 42 | 1.80M |

Table 2: EA-CKG statistics. Reddit and Twitter are two conversation datasets used in our experiments.

**Latent Concepts Construction** During training and inference, given a message $X_i$ and a desired emotion $e_i$, we construct the latent concepts of the response based on EA-CKG embeddings. Specifically, we first train a well-established knowledge embedding model, i.e., TransE (Bordes et al. 2013)[5], on the entire EA-CKG to learn global concept and relation embeddings. The embeddings in TransE are learned such that the score $-||\mathbf{h}+\mathbf{r}-\mathbf{t}||_2$ for a correct triplet $(h, r, t)$ is much higher than a corrupted one, where $\mathbf{h}, \mathbf{r}, \mathbf{t}$ denote the TransE embeddings of $h, r, t$, respectively, and $||\mathbf{h}||_2 = 1$ and $||\mathbf{t}||_2 = 1$ (Bordes et al. 2013). Hence, given a message concept $h$, a relation $r$ and a response concept $t$, we can estimate the relatedness between $h$ and $t$ via $r$ as follows:

$$\text{score}(h, r, t) = (\mathbf{h} + \mathbf{r})^\top \mathbf{t}. \quad (2)$$

We then obtain the top $m$ related latent concepts of the response from EA-CKG, i.e., $\{t\}_1^m$, as follows:

$$\{t_i\}_1^m = \underset{t}{\text{top}}(\text{score}(h, r, t)), \quad (3)$$

where $h \in C_{X_i}$, $r \in R \cup \{e_i\}$, $C_{X_i}$ denotes all concepts in $X_i$, $R$ denotes all 36 relations in ConceptNet, and $t$ is searched over the concept vocabulary of EA-CKG. For messages without any concepts[6], we use a null message concept whose embedding is the average of all concept embeddings.

**Framework Analysis** Our framework constructs plausible relational ($r \in R$) and emotional ($r = e_i$) concepts for the response. By leveraging the EA-CKG embeddings, our

[2]We remove non-English and rare concepts.

[3]We consider concept pairs whose frequency $\geq 5$ and PPMI $\geq$ 1 as strongly associated pairs (CCP).

[4]We associate a CCP $\{w_1, w_2\}$ with emotion $e$ if $\text{PPMI}(\{w_1, w_2\}, e) - \max_{e_i \neq e} \text{PPMI}(\{w_1, w_2\}, e_i) \geq 1$.

[5]We adopt TransE because it achieves only marginally worse performance than RotatE (Sun et al. 2019), a state-of-the-art knowledge graph embedding model, for triplet classification on ConceptNet, but much faster in inference.

[6]Around 3% messages do not have any concepts.

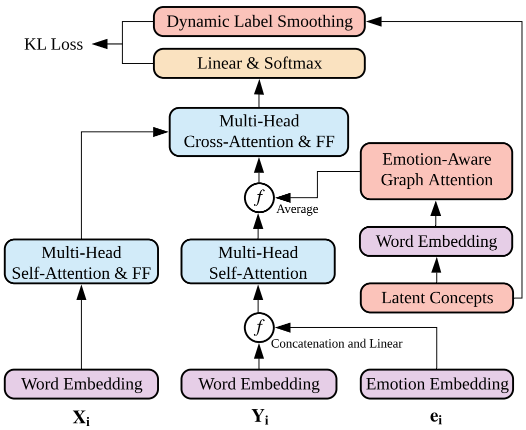


Figure 2: Architecture of our Transformer-based conversational model. The positional encoding, residual connection, and layer normalization are omitted in the illustration for brevity.

framework inherits the ideas from knowledge base completion and has two major advantages over the graph search methods used in existing models (Zhou et al. 2018b; Zhang et al. 2020) to find related concepts: 1) our framework can find concepts that are both commonsense-aware and emotional due to the incorporation of emotional triplets in EA-CKG, e.g., *tasty* is found given *beer* and *joy* whereas *bland* is found given *beer* and *sadness*; and 2) our framework can not only traverse through the EA-CKG to find related concepts in a multi-hop neighborhood but also discover an arbitrary number of novel related concepts using Equation 3, without being limited by the CKG coverage (see Result Analysis).

## Incorporating Latent Concepts

After obtaining the latent concepts, we propose three methods to collaboratively incorporate them into our Transformer-based conversational model (Vaswani et al. 2017), as illustrated in Figure 2. Note that similar to the idea of persona embedding (Li et al. 2016b), we additionally employ an emotion embedding layer in our decoder.

**Emotion-Aware Graph Attention** We incorporate latent concepts into the decoder using an emotion-aware graph attention (**EAGA**) prior to the cross-attention layer, inspired by (Zhong, Wang, and Miao 2019b). We assume that important latent concepts are those related to the message concepts and have strong emotional intensity. The relatedness between concepts is obtained from Equation 2. The emotional intensity of a concept is computed based on an emotion lexicon NRC_VAD (Mohammad 2018) and an emotional intensity computation method (Zhong, Wang, and Miao 2019b). We expand the size of NRC_VAD from 20K to 34K using synonym expansion for better coverage[7].

Formally, let $\{t_1, t_2, ..., t_m\}$ be the latent concepts of response $Y_i$ obtained from Equation 3, $\{s_1, s_2, ..., s_m\}$ be their relatedness scores obtained from Equation 2, and $\{q_1, q_2, ..., q_m\}$ be their emotion intensities based on NRC_VAD, we compute the latent concept embedding of $Y_i$, i.e., $\mathbf{C}_{Y_i}$, as follows:

$$\mathbf{C}_{Y_i} = \sum_{i=1}^{m} \beta_i \mathbf{t}_i, \quad (4)$$

where $\mathbf{t}_i$ denotes the word embedding of $t_i$ and $\beta_i$ is computed as follows:

$$\beta_i = \lambda_i \frac{\exp(\delta_{1i} s_i)}{\sum_j \exp(\delta_{1j} s_j)} + (1 - \lambda_i) \frac{\exp(\delta_{2i} q_i)}{\sum_j \exp(\delta_{2j} q_j)}, \quad (5)$$

where $\lambda_i$ denotes the trade-off coefficient between relatedness and emotional intensity, and $\delta_{1i}, \delta_{2i}$ denote softmax temperatures. Note that $\lambda_i$, $\delta_{1i}$ and $\delta_{2i}$ are concept-specific and can be fixed *a prior* or learned during training. The obtained latent concept embedding $\mathbf{C}_{Y_i}$ is then averaged with the response representation prior to being fed to the cross-attention layer. Compared with the graph attention in CCM (Zhou et al. 2018b), EAGA measures concept relatedness using translation-based distance in TransE instead of MLP and additionally considers the emotion property of concepts.

**Dynamic Label Smoothing** Label smoothing is conventionally adopted in the Transformer (Vaswani et al. 2017) to improve translation quality. We propose a simple but effective dynamic label smoothing (**DLS**) method to explicitly enforce the supervision of latent concepts in producing concept-related responses, as well as to stabilize the learning process. Specifically, starting from the conventional label smoothing, we linearly increase the smoothing values for latent concepts with the training step and decrease the smoothing values for other words in the vocabulary. Note that the smoothing value of the target word remains unchanged. The maximum of the total smoothing value for latent concepts is a hyper-parameter to be tuned in experiments. We optimize model parameters to minimize the Kullback-Leibler (KL) loss (Kullback and Leibler 1951).

**Concept-Aware Top-$K$ Decoding** During inference, we propose a concept-aware top-$K$ decoding (**CATD**) method to encourage the generation of words that are more related to the associated latent concepts. Formally, given the conventional top-$K$ unnormalized token probabilities $P(w_1), ..., P(w_k)$, our concept-aware token probability $P^{'}$ for $w_i, i = 1, ..., k$, is computed as follows:

$$P^{'}(w_i) = P(w_i) * P_c^{\gamma}(w_i), \quad (6)$$

where $\gamma$ denotes a trade-off hyper-parameter between fluency and relatedness, and $P_c(w_i)$ is computed as follows:

$$P_c(w_i) = \frac{\exp(\mathbf{C}_Y^{\top} \mathbf{w}_i)}{\sum_{i=1}^{k} \exp(\mathbf{C}_Y^{\top} \mathbf{w}_i)}, \quad (7)$$

where $\mathbf{C}_Y$ denotes the latent concept embedding obtained from Equation 4 during inference. One merit of CATD is that it only reorders top-$K$ tokens by additionally considering their relatedness to latent concepts and thus does not introduce unlikely tokens into the sampling process.

[7]The expanded NRC_VAD covers more than 97% tokens in the datasets used in our experiments.

| | | Reddit | Twitter |
|---|---|---|---|
| Training | Neutral | 268K | 649K |
| | Joy | 232K | 308K |
| | Sadness | 236K | 302K |
| | Surprise | 551K | 543K |
| | Fear | 156K | 325K |
| | Anger | 132K | 373K |
| | Total | 1.58M | 2.50M |
| Validation | Total | 49K | 50K |
| Testing | Total | 49K | 50K |

Table 3: Dataset statistics.

# Experimental Settings

In this section, we present the datasets, evaluation metrics, baselines, and model settings.

## Datasets

We conduct experiments on two large-scale datasets, namely Reddit and Twitter. The Reddit dataset is obtained from comments on the CasualConversation subreddit[8] discussing a variety of casual topics[9]. The Twitter dataset is obtained from chats on twitter.com[10]. We truncate each sentence to a maximum of 30 tokens and use the most frequent 30K tokens as the vocabulary for each dataset.

To obtain the ground-truth emotion label for each response, similar to (Zhou et al. 2018a; Song et al. 2019), we train an emotion classifier on emotional conversations. Specifically, we use the emotional tweets (Mohammad 2012; Mohammad et al. 2018) to train the classifier. We consider neutral and Ekman's six basic emotions (Ekman 1992): joy, sadness, surprise, fear, and anger, but exclude disgust due to its small amount of training samples in the emotional tweets. We propose an emotion classifier based on DeepMoji embeddings (Felbo et al. 2017) followed by a linear layer and a softmax layer. Our classifier achieves an accuracy of 0.562 on a balanced test dataset, outperforming several competitive baselines such as BiLSTM (0.446), CNN (0.547), BERT (Devlin et al. 2019) (0.530) and XLNet (Yang et al. 2019) (0.522). We then use the trained emotion classifier to annotate the responses in the datasets. The statistics of the annotated datasets are presented in Table 3.

## Evaluation Metrics

We conduct both automatic and human evaluations. Automatic evaluation metrics include 1) **Fluency**: perplexity (PPL), which measures the confidence of the generated responses; 2) **Diversity**: distinct-1 (dist-1) and distinct-2 (dist-2) (Li et al. 2016a), which measure the percentage of unique unigrams and bigrams in the generated responses, respectively; 3) **Emotion Accuracy (EA)**: the emotion accuracy of the generated responses measured by our trained emotion classifier; and 4) **Commonsense Awareness (CA)**: the average number of commonsense triplets in one pair of message and generated response, measured by ConceptNet.

[8]https://www.reddit.com/r/CasualConversation/

[9]https://files.pushshift.io/reddit/comments/

[10]https://github.com/Marsan-Ma/chat corpus/

Following (Zhou et al. 2018a), we conduct human evaluations to measure both **content quality** (rating scale in $\{0, 1, 2\}$) and **emotion quality** (rating scale in $\{0, 1\}$) of the generated responses. Content quality measures whether the response is natural and related to the message, as well as how commonsense-aware the response is. Emotion quality measures whether the response expresses the desired emotion appropriately and accurately. We randomly sample 200 test messages and emotions to generate 200 responses for each model. Each response is evaluated by three annotators.

## Baselines

We compare CARE with the following baselines:
**Vanilla Models**: Seq2Seq (Vinyals and Le 2015) and Transformer (Vaswani et al. 2017).
**Commonsense-Aware Models**: CCM (Zhou et al. 2018b) and ConceptFlow (Zhang et al. 2020). ConceptFlow leverages multi-hop knowledge triplets and is a state-of-the-art model for commonsense-aware response generation.
**Emotional Models**: ECM (Zhou et al. 2018a) and EmoDS (Song et al. 2019). EmoDS is a state-of-the-art model for emotional response generation.
**Pre-trained Model**: CTRL (Keskar et al. 2019). CTRL is a large pre-trained conditional language model with 1.6 billion parameters trained on 140GB of text. We fine-tune CTRL on our training conversations such that it is able to produce emotional responses. CTRL has also been shown to contain commonsense knowledge (Petroni et al. 2019; Bosselut et al. 2019).

## Model Settings

We use the same hyper-parameters for both datasets. Our TransE embeddings have a dimension of 100 and achieve an accuracy of 0.89 for triplet classification on EA-CKG. Our Transformer model has 1 layer and 4 attention heads. We initialize the word embedding layer with pre-trained GloVe embeddings (Pennington, Socher, and Manning 2014) of size 300. The emotion embedding and feedforward layers have sizes of 50 and 512, respectively. We train our model using Adam (Kingma and Ba 2014) with learning rate of 1, batch size of 64, and dropout of 0.1 for 80K steps, including 6K steps for warmup. We empirically construct 30 relational latent concepts and 10 emotional latent concepts for each response using Equation 3. We use label smoothing of 0.1, total smoothing value of 0.08 for latent concepts in DLS, and top-10 decoding with $\gamma = 1$ in CATD.

# Result Analysis

In this section, we discuss our evaluation results, model analysis, case study, error analysis and limitation.

## Comparison with Baselines

We present the results of automatic evaluations in Table 4. Seq2Seq achieves the lowest perplexity while Transformer achieves slightly better diversity than Seq2Seq. Commonsense-aware models, i.e., CCM and ConceptFlow, obtain slightly better diversity and CA; however, they are

| Models | Reddit | | | | | Twitter | | | | | Size | IT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PPL | Dist-1 | Dist-2 | EA | CA | PPL | Dist-1 | Dist-2 | EA | CA | | |
| Seq2Seq | **57.2** | 0.0035 | 0.0347 | - | 0.1349 | **79.7** | 0.0047 | 0.0522 | - | 0.1653 | 38M | **1.0x** |
| Transformer | 63.8 | 0.0032 | 0.0371 | - | 0.1224 | 90.1 | 0.0053 | 0.0563 | - | 0.1728 | **20M** | 1.5x |
| CCM | 62.3 | 0.0046 | 0.0469 | - | 0.1222 | 82.5 | 0.0060 | 0.0663 | - | 0.1835 | 74M | 5.9x |
| ConceptFlow | 60.1 | 0.0047 | 0.0458 | - | 0.1375 | 89.1 | 0.0051 | 0.0556 | - | 0.1893 | 33M | 21.8x |
| ECM | 65.6 | 0.0044 | **0.0506** | 0.5893 | 0.1105 | 91.3 | 0.0056 | 0.0630 | 0.5619 | 0.1650 | 40M | 2.0x |
| EmoDS | 76.6 | 0.0030 | 0.0455 | 0.6186 | 0.1107 | 113.5 | 0.0030 | 0.0450 | 0.5950 | 0.1599 | 46M | 1.5x |
| CTRL | - | **0.0068** | 0.0447 | 0.3425 | 0.1502 | - | **0.0108** | **0.0851** | 0.3995 | 0.1958 | 1.6B | 1876.7x |
| Ours (CARE) | 70.4 | 0.0049 | 0.0460 | **0.6840** | **0.1538** | 100.1 | 0.0064 | 0.0775 | **0.6693** | **0.2304** | 20M | 1.9x |

Table 4: Automatic evaluation results. Size denotes model size. IT denotes inference time relative to Seq2Seq.

| | Models | Neutral | | Joy | | Sadness | | Surprise | | Fear | | Anger | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cont | Emot | Cont | Emot | Cont | Emot | Cont | Emot | Cont | Emot | Cont | Emot | Cont | Emot |
| **Reddit** | Seq2Seq | 0.62 | 0.34 | 0.79 | 0.32 | 0.69 | 0.15 | 0.78 | 0.35 | **0.72** | 0.19 | 0.74 | 0.08 | 0.73 | 0.24 |
| | ConceptFlow | 0.82 | 0.45 | 0.96 | 0.35 | 0.81 | 0.17 | 0.89 | 0.31 | 0.70 | 0.16 | 0.76 | 0.15 | 0.83 | 0.26 |
| | EmoDS | 0.76 | 0.66 | 0.89 | 0.72 | 0.86 | **0.67** | 0.71 | 0.52 | 0.63 | 0.41 | 0.68 | 0.38 | 0.75 | 0.56 |
| | CTRL | **0.92** | 0.50 | **1.08** | 0.63 | **1.03** | 0.42 | 0.79 | 0.34 | 0.66 | 0.24 | **0.93** | 0.38 | **0.90** | 0.42 |
| | Ours (CARE) | 0.78 | **0.68** | 0.98 | **0.75** | 0.88 | 0.63 | **0.92** | **0.76** | 0.63 | **0.44** | 0.81 | **0.42** | 0.84 | **0.62** |
| **Twitter** | Seq2Seq | 0.92 | 0.33 | 0.76 | 0.23 | 0.79 | 0.21 | 0.85 | 0.17 | 0.81 | 0.25 | 0.99 | 0.29 | 0.86 | 0.25 |
| | ConceptFlow | 0.97 | 0.42 | 0.91 | 0.28 | 0.98 | 0.22 | 1.03 | 0.19 | 0.87 | 0.21 | 0.85 | 0.26 | 0.93 | 0.27 |
| | EmoDS | 0.82 | 0.46 | 0.78 | 0.48 | 0.91 | 0.56 | 0.93 | 0.63 | 0.79 | 0.65 | 0.84 | **0.65** | 0.84 | 0.57 |
| | CTRL | **1.08** | 0.54 | **1.05** | **0.62** | **1.16** | 0.50 | **1.21** | 0.68 | 0.92 | 0.71 | **1.12** | 0.61 | **1.09** | 0.61 |
| | Ours (CARE) | 0.87 | **0.57** | 0.83 | 0.58 | 1.13 | **0.62** | 1.15 | **0.71** | **0.93** | **0.74** | 0.94 | 0.63 | 0.96 | **0.64** |

Table 5: Human evaluation results. Cont and Emot denote content quality and emotion quality, respectively. The inter-annotator agreement, measured by Fleiss' Kappa (Fleiss and Cohen 1973), are 0.441 and 0.626 for content and emotion on Reddit, respectively, and 0.479 and 0.673 for content and emotion on Twitter, respectively. Both datasets obtain "moderate agreement" and "substantial agreement" for content and emotion, respectively.

unable to generate responses with specified emotions. Emotional models, i.e., ECM and EmoDS, achieve the highest EA among all baselines but the worst in perplexity and CA, suggesting that they only specialize in emotion and neglect commonsense. CTRL achieves the highest diversity among all models, partially due to its large vocabulary size of 250K. However, it obtains an inferior EA. Our model achieves better EA and CA than all baselines, including CTRL, which is also capable of producing commonsense-aware emotional responses.

We present the results of human evaluations in Table 5. The responses of non-emotional models are generated via top-10 decoding six times. ConceptFlow obtains similar emotion quality but noticeably better content quality than Seq2Seq due to its incorporation of multi-hop triplets. EmoDS achieves comparable content quality but much better emotion quality than Seq2Seq. CTRL obtains the best content quality among all models but only mediocre emotion quality, especially on Reddit. Our model performs best in emotion quality ($t$-test, $p < 0.01$). In addition, our model achieves significantly better content quality than EmoDS ($t$-test, $p < 0.01$), showing that our model can produce better commonsense-aware emotional responses than EmoDS. Finally, our model outperforms ConceptFlow, a competitive commonsense-aware model, in content quality, possibly because the graph search method in ConceptFlow heavily relies on the coverage of ConceptNet to extract knowledge triplets, but ConceptNet only has an average coverage of 27% on Reddit and Twitter. In contrast, our model has less

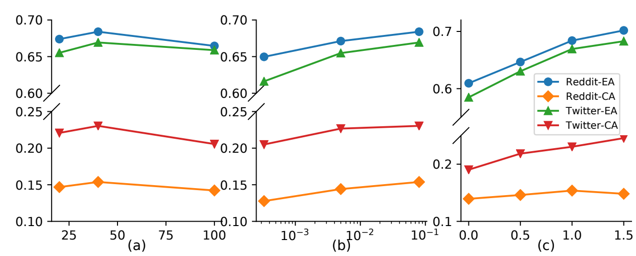


Figure 3: Hyper-parameter analysis on EA and CA. (a) Different number of latent concepts for each response (see $m$ in Equation 3), where 1/4 latent concepts are emotional. (b) Different total smoothing values for latent concepts in DLS. (c) Different $\gamma$ (see Equation 6) in CATD.

such restriction and can construct an arbitrary number of latent concepts given any input message.

We report model complexity in the rightmost columns of Table 4. Our model has comparable space and time complexity with vanilla baselines. In contrast, CTRL is around 80x larger and 1,000x slower than our model, rendering it intractable for real-time applications.

## Model Analysis

We conduct ablation study, as shown in Table 6. Removing any component except EAGA from our model leads to much worse performance in both EA and CA. In particular, we

| | Reddit | | | | | Twitter | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Models** | PPL | Dist-1 | Dist-2 | EA | CA | PPL | Dist-1 | Dist-2 | EA | CA |
| Ours (CARE) | **70.4** | **0.0049** | 0.0460 | **0.6840** | 0.1538 | **100.1** | 0.0064 | 0.0775 | **0.6693** | 0.2304 |
| -ET+EL | 72.2 | 0.0040 | 0.0428 | 0.6518 | 0.1332 | 100.8 | 0.0057 | 0.0669 | 0.6266 | 0.2077 |
| -TransE | 72.8 | 0.0039 | 0.0430 | 0.6595 | 0.1261 | 101.8 | 0.0057 | 0.0660 | 0.6391 | 0.1960 |
| -EAGA | 79.6 | 0.0045 | **0.0484** | 0.6258 | **0.1635** | 116.3 | **0.0080** | **0.1303** | 0.4775 | **0.3512** |
| -DLS | 72.6 | 0.0038 | 0.0441 | 0.6497 | 0.1277 | 100.7 | 0.0056 | 0.0682 | 0.6162 | 0.2050 |
| -DLS + LS | 72.5 | 0.0040 | 0.0443 | 0.6421 | 0.1318 | 101.2 | 0.0055 | 0.0675 | 0.6194 | 0.2013 |
| -CATD | **70.4** | 0.0036 | 0.0373 | 0.6094 | 0.1394 | **100.1** | 0.0059 | 0.0630 | 0.5848 | 0.1903 |

Table 6: Ablation study. -ET+EL: replace the tails of the extracted emotional triplets (ET) by randomly sampled corresponding emotional words from an emotion lexicon (EL) (Mohammad and Turney 2013). -TransE: instead of using TransE, search neighbors with a growing neighborhood size (up to 3) on EA-CKG to find latent concepts. -EAGA: remove the emotion-aware graph attention. -DLS: remove the dynamic label smoothing. -DLS+LS: replace the dynamic label smoothing by conventional label smoothing (LS) of 0.1. -CATD: replace the concept-aware top-$K$ decoding by the conventional top-$K$ decoding.

| | | Why do you avoid starbucks? | I'd like to confess that I'm a cat person way more then a dog person. |
|---|---|---|---|
| Seq2Seq | | It's not bad. | I hope you get the chance. |
| Ours (CARE) | Neutral | It's just a general dislike of the **company**. | I'm curious to hear the story behind this! |
| | Joy | **Tea** is my *favorite*. | Yes! I *love* their **eyes**, I *love* them. |
| | Sadness | I'm not a big fan of **coffee**. It 's too *bitter* for me. | I don't understand the difference though :( |
| | Surprise | I'm not sure. Maybe because of the **coffee shop** thing? | What kind of **cat** you like ? |
| | Fear | I'm not sure. It's just a *weird* feeling. | That's the *worst*! |
| | Anger | I hate the taste of **coffee**. It's just so *bitter* and *gross*. | That's a little *annoying*! |

Table 7: Case studies. Words in bold and italic denote relational and emotional latent concepts, respectively.

observe that 1) our approach of constructing latent concepts performs better than alternatives (-ET+EL and -TransE); and 2) the removal of EAGA leads to significantly higher perplexity, diversity, and CA. The higher perplexity may be attributed to the additional supervisions of DLS on latent concepts, which are not explicitly incorporated into the model due to the lack of EAGA. The higher diversity and CA may be attributed to the untrained $\lambda$, $\delta_1$, and $\delta_2$ (see Equation 5), which sometimes leads to ungrammatical but diverse latent concepts during decoding. Our observation validates the importance of EAGA in attending more related latent concepts.

We analyze the impact of model hyper-parameters on EA and CA, as shown in Figure 3. Using $m = 40$ latent concepts achieves the sweet spot for model complexity. Regarding DLS, increasing the total smoothing values for latent concepts in the $[0, 0.08]$ range improves model performance. However, we do observe degraded fluency when using larger smoothing values, which is expected because the true learning signal is weakened. Increasing $\gamma$ in CATD consistently improves EA and CA for our model. However, models with larger $\gamma$, e.g., 1.5, sometimes produce unfluent long responses due to its overemphasizes on latent concepts.

## Case Study and Error Analysis

We present two sample cases in Table 7. Given a message and desired emotions, our model produces commonsense-aware responses with the desired emotions, guided by both relational and emotional latent concepts. For example, given "starbucks" and anger, the relational latent concept "coffee" and emotional latent concept "gross" are constructed and incorporated into response generation. However, we do observe bad cases where the latent concepts overemphasize on emotional intensity, and the response becomes unnatural.

## Limitation

One major limitation of our work is the mediocre accuracy of our trained emotion classifier, which can be attributed to the unavailability of large-scale datasets for emotional conversations and sentences. Nevertheless, our proposed lightweight classifier obtains better performance than the best models reported in (Zhou et al. 2018a; Song et al. 2019) and BERT. A potential solution to this limitation is to leverage few-shot learning on BERT-like models.

# Conclusion

We propose CARE as the first attempt to test the hypothesis that combing rationality (commonsense) and emotion into conversational agents can improve response quality and human ratings. Specifically, we build an EA-CKG and leverage its TransE embeddings to allow CARE to reason over the EA-CKG and construct both relational and emotional latent concepts. We further propose three methods to collaboratively incorporate the latent concepts into response generation. Extensive ablation studies show that our methods of constructing and incorporating latent concepts outperform alternative methods. In addition, both automatic and human evaluations show that CARE can produce more accurate and commonsense-aware emotional responses than state-of-the-art commonsense-aware models and emotional models. Finally, our work provides empirical evidence for our hypothesis. In the future, we plan to extend our work to other aspects of rationality, e.g., logical reasoning.

## Acknowledgments

This research is supported, in part, by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI) (Alibaba-NTU-AIR2019B1), Nanyang Technological University, Singapore. This research is also supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003) and under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is also supported, in part, by the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017 and MOH/NIC/HAIG03/2017).

## References

Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; and Mou, L. 2018. Affective Neural Response Generation. In *ECIR*, 154–166.

Banerjee, P.; Pal, K. K.; Mitra, A.; and Baral, C. 2019. Careful Selection of Knowledge to Solve Open Book Question Answering. In *ACL*, 6120–6129.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, 2787–2795.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL*, 4762–4779.

Church, K. W.; and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22–29.

Colman, A. M. 2003. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences* 26(2): 139–153.

De Sousa, R. 1990. *The rationality of emotion*. MIT Press.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion* 6(3-4): 169–200.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*, 1615–1625.

Fleiss, J. L.; and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33(3): 613–619.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *AAAI*, 5110–5117.

Ghosh, S.; Chollet, M.; Laksana, E.; Morency, L.-P.; and Scherer, S. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *ACL*, 634–642.

Han, S.; Bang, J.; Ryu, S.; and Lee, G. G. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *SIGDIAL*, 129–133.

Hasegawa, T.; Kaji, N.; Yoshinaga, N.; and Toyoda, M. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 964–972.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *ICML*, 1587–1596.

Keltner, D.; and Haidt, J. 1999. Social functions of emotions at four levels of analysis. *Cognition & Emotion* 13(5): 505–521.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* .

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1): 79–86.

Levy, O.; Goldberg, Y.; and Dagan, I. 2015. Improving distributional similarity with lessons learned from word embeddings. *TACL* 3: 211–225.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL*, 110–119.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A Persona-Based Neural Conversation Model. In *ACL*, 994–1003.

Li, J.; and Sun, X. 2018. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. *arXiv preprint arXiv:1806.07000* .

Li, P.; and Tuzhilin, A. 2019. Towards Controllable and Personalized Review Generation. In *EMNLP-IJCNLP*, 3228–3236.

Li, X.; Taheri, A.; Tu, L.; and Gimpel, K. 2016c. Commonsense knowledge base completion. In *ACL*, 1445–1455.

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 2019. MoEL: Mixture of Empathetic Listeners. In *EMNLP-IJCNLP*, 121–132.

Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018. Knowledge diffusion for neural dialogue generation. In *ACL*, 1489–1498.

Madotto, A.; Wu, C.-S.; and Fung, P. 2018. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. In *ACL*, 1468–1478.

Mohammad, S. 2012. #Emotional Tweets. In *SemEval*, 246–255.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *ACL*, 174–184.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *SemEval*, 1–17.

Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29(3): 436–465.

Moon, S.; Shah, P.; Kumar, A.; and Subba, R. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *ACL*, 845–854.

Peng, Y.; Fang, Y.; Xie, Z.; and Zhou, G. 2019. Topic-enhanced emotional conversation generation with attention mechanism. *Knowledge-Based Systems* 163: 429–437.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *EMNLP-IJCNLP*, 2463–2473.

Pham, M. T. 2007. Emotion and rationality: A critical review and interpretation of empirical evidence. *Review of General Psychology* 11(2): 155–178.

Prendinger, H.; and Ishizuka, M. 2005. The empathic companion: A character-based interface that addresses users'affective states. *Applied Artificial Intelligence* 19(3-4): 267–285.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL*, 5370–5381.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637* .

Ross, J. J. 1978. Rationality and common sense. *Philosophy* 53(205): 374–381.

Saito, I.; Nishida, K.; Asano, H.; and Tomita, J. 2018. Commonsense knowledge base completion and generation. In *CoNLL*, 141–150.

Song, Z.; Zheng, X.; Liu, L.; Xu, M.; and Huang, X.-J. 2019. Generating Responses with a Specific Emotion in Dialog. In *ACL*, 3685–3695.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. In *AAAI*, 4444–4451.

Sun, H.; Dhingra, B.; Zaheer, M.; Mazaitis, K.; Salakhutdinov, R.; and Cohen, W. 2018. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. In *EMNLP*, 4231–4242.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*.

Tuan, Y.-L.; Chen, Y.-N.; and Lee, H.-y. 2019. DyKgChat: Benchmarking Dialogue Generation Grounding on Dynamic Knowledge Graphs. In *EMNLP-IJCNLP*, 1855–1865.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Vinyals, O.; and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* .

Wu, C.-S.; Socher, R.; and Xiong, C. 2019. Global-to-local Memory Pointer Networks for Task-Oriented Dialogue. In *ICLR*.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic aware neural response generation. In *AAAI*, 3351–3357.

Xu, C.; Wu, W.; Tao, C.; Hu, H.; Schuerman, M.; and Wang, Y. 2019. Neural Response Generation with Meta-words. In *ACL*, 5416–5426.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *AAAI*, 4970–4977.

Zhang, H.; Liu, Z.; Xiong, C.; and Liu, Z. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In *ACL*.

Zhong, P.; Wang, D.; and Miao, C. 2019a. An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss. In *AAAI*, 7492–7500.

Zhong, P.; Wang, D.; and Miao, C. 2019b. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *EMNLP-IJCNLP*, 165–176.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018a. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *AAAI*, 730–739.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018b. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*, 4623–4629.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2018c. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics* 0: 1–62.

Zhou, X.; and Wang, W. Y. 2018. MojiTalk: Generating Emotional Responses at Scale. In *ACL*, 1128–1137.