An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss

Peixiang Zhong,[1,2] **Di Wang,**[1] **Chunyan Miao**[1,2,3]
[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly
[2]Alibaba-NTU Singapore Joint Research Institute
[3]School of Computer Science and Engineering
Nanyang Technological University, Singapore
peixiang001@e.ntu.edu.sg, {wangdi, ascymiao}@ntu.edu.sg

Abstract

Affect conveys important implicit information in human communication. Having the capability to correctly express affect during human-machine conversations is one of the major milestones in artificial intelligence. In recent years, extensive research on open-domain neural conversational models has been conducted. However, embedding affect into such models is still under explored. In this paper, we propose an end-to-end affect-rich open-domain neural conversational model that produces responses not only appropriate in syntax and semantics, but also with rich affect. Our model extends the Seq2Seq model and adopts VAD (Valence, Arousal and Dominance) affective notations to embed each word with affects. In addition, our model considers the effect of negators and intensifiers via a novel affective attention mechanism, which biases attention towards affect-rich words in input sentences. Lastly, we train our model with an affect-incorporated objective function to encourage the generation of affect-rich words in the output responses. Evaluations based on both perplexity and human evaluations show that our model outperforms the state-of-the-art baseline model of comparable size in producing natural and affect-rich responses.

Introduction

Affect is a psychological experience of feeling or emotion. As a vital part of human intelligence, having the capability to recognize, understand and express affect and emotions like human has been arguably one of the major milestones in artificial intelligence (Picard 1997).

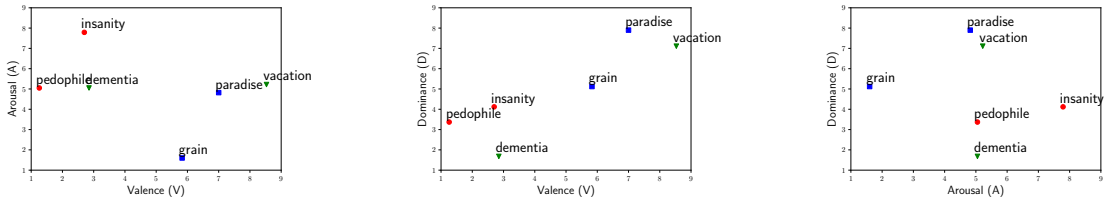
Open-domain conversational models aim to generate coherent and meaningful responses when given user input sentences. In recent years, neural network based generative conversational models relying on Sequence-to-Sequence network (Seq2Seq) (Sutskever, Vinyals, and Le 2014) have been widely adopted due to its success in neural machine translation. Seq2Seq based conversational models have the advantages of end-to-end training paradigm and unrestricted response space over conventional retrieval-based models. To make neural conversational models more engaging, various techniques have been proposed, such as using stochastic latent variable (Serban et al. 2017) to promote response diversity and encoding topic (Xing et al. 2017) into conversational models to produce more coherent responses.

However, embedding affect into neural conversational models has been seldom explored, despite that it has many benefits such as improving user satisfaction (Callejas, Griol, and López-Cózar 2011), fewer breakdowns (Martinovski and Traum 2003), and more engaged conversations (Robison, McQuiggan, and Lester 2009). For real-world applications, Fitzpatrick, Darcy, and Vierhile (2017) developed a rule-based empathic chatbot to deliver cognitive behavior therapy to young adults with depression and anxiety, and obtained significant results on depression reduction. Despite of these benefits, there are a few challenges in the affect embedding in neural conversational models that existing approaches fail to address: (i) It is difficult to capture the emotion of a sentence, partly because negators and intensifiers often change its polarity and strength. Handling negators and intensifiers properly still remains as a challenge in sentiment analysis. (ii) It is difficult to embed emotions naturally in responses with correct grammar and semantics (Ghosh et al. 2017).

In this paper, we propose an end-to-end single-turn open-domain neural conversational model to address the aforementioned challenges to produce responses that are natural and affect-rich. Our model extends Seq2Seq model with attention (Luong, Pham, and Manning 2015). We leverage an external corpus (Warriner, Kuperman, and Brysbaert 2013) to provide affect knowledge for each word in the Valence, Arousal and Dominance (VAD) dimensions (Mehrabian 1996). We then incorporate the affect knowledge into the embedding layer of our model. VAD notation has been widely used as a dimensional representation of human emotions in psychology and various computational models, e.g., (Wang, Tan, and Miao 2016; Tang et al. 2017). 2D plots of selected words with extreme VAD values are shown in Figure 1. To capture the effect of negators and intensifiers, we propose a novel biased attention mechanism that explicitly considers negators and intensifiers in attention computation. To maintain correct grammar and semantics, we train our Seq2Seq model with a weighted cross-entropy loss that encourages the generation of affect-rich words without degrading language fluency.

Our main contributions are summarized as follows:

- For the first time, we propose a novel affective attention mechanism to incorporate the effect of negators and intensifiers in conversation modeling. Our mechanism in-



(a) V-A ratings (b) V-D ratings (c) A-D ratings

Figure 1: 2D plot of words with either highest or lowest ratings in valence (V), arousal (A) or dominance (D) in the corpus.

troduces only a small number of additional parameters.

- For the first time, we apply weighted cross-entropy loss in conversation modeling. Our affect-incorporated weights achieve a good balance between language fluency and emotion quality in model responses. Our empirical study does not show performance degradation in language fluency while producing affect-rich words.
- Overall, we propose *Affect-Rich Seq2Seq* (AR-S2S), a novel end-to-end affect-rich open-domain neural conversational model incorporating external affect knowledge. Human preference test shows that our model is preferred over the state-of-the-art baseline model in terms of both content quality and emotion quality by a large margin.

Related Work

Prior studies on affective conversational systems mainly focused on rule-based systems, which require an extensive hand-crafted rule base. For example, Ochs, Pelachaud, and Sadek (2008) designed an empathetic virtual agent that can express emotions based on cognitive appraisal theories (Hewstone and Stroebe 2001), which require numerous event-handling rules to be implemented. Another example is the Affect Listeners (Skowron 2010), which are conversational systems aiming to detect and adapt to the affective states of users. However, their detection and adaptation mechanisms heavily rely on hand-crafted features such as letter capitalization, punctuation and emoticons.

In recent years, there is an emerging research trend in end-to-end neural network based generative conversational systems (Vinyals and Le 2015; Shang, Lu, and Li 2015). To improve the content quality of neural conversational models, many techniques have been proposed, such as improving response diversity using Conditional Variational Autoencoders (CVAE) (Zhao, Zhao, and Eskenazi 2017) and encoding commonsense knowledge using external facts corpus (Ghazvininejad et al. 2018).

However, few work investigated the problems in improving the emotion quality of neural conversational models. Emotional Chatting Machine (ECM) (Zhou et al. 2018) is a Seq2Seq conversational model that generates responses with user-input emotions. It employs an internal memory module to model implicit emotional changes and an external memory module to help generate more explicit emotional words. The main objective of ECM is to produce responses according to explicit user-input emotions. While our model focuses on enriching affect in generated responses. Similar to ECM, Mojitalk (Zhou and Wang 2018) presents a few generative models, including Seq2Seq, CVAE and Reinforced CVAE, to generate responses according to explicit user-input emojis. Both ECM and Mojitalk do not consider emotions in input sentences when generating emotional responses. In comparison, our model considers them naturally with focuses on affect-rich words and avoids an additional step of determining which emotion to respond with during conversations. Asghar et al. (2018) introduces a Seq2Seq model with three extensions to incorporate affects into conversations. Similar to their work, we also adopt the approach of using VAD embedding to encode affects. However, we perform extra preprocessing on VAD embedding to improve model performance. In addition, we specifically consider the effect of negators and intensifiers via a novel affective attention mechanism when generating affect-rich responses.

Seq2Seq with Attention

Prior to introducing our proposed model, we briefly describe the vanilla Seq2Seq model with attention. Seq2Seq model is a neural network model mapping the input sequence to the output sequence. Specifically, it uses a Recurrent Neural Network (RNN) encoder to encode the variable length input sequence $X = (x_1, x_2, ..., x_T)$ as a vector of fixed dimensionality $\mathbf{h_T}$ and an RNN decoder to decode $\mathbf{h_T}$ as the variable length output sequence $Y = (y_1, y_2, ..., y_{T'})$. The objective function of Seq2Seq is to maximize

$$p(Y|X) = p(y_1|\mathbf{h}_T) \prod_{t'=2}^{T'} p(y_{t'}|\mathbf{h}_T, y_1, ..., y_{t'-1}),$$
$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, x_t), \forall t = 1, 2, ..., T, \quad (1)$$

where $\mathbf{h_t}$ denotes the hidden state of input sequence at time step t and $\mathbf{h_0}$ is usually initialized as a zero vector. Function f denotes a non-linear transformation, which usually takes the form of recurrent models such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) or Gated Recurrent Units (GRU) (Cho et al. 2014).

After encoding X as $\mathbf{h_T}$, the decoder updates its decoder hidden state $\mathbf{s}_{t'}$ by taking the previous hidden state $\mathbf{s}_{t'-1}$ and previous output $y_{t'-1}$ as inputs:

$$\mathbf{s}_{t'} = g(\mathbf{s}_{t'-1}, y_{t'-1}), \forall t' = 1, 2, ..., T', \quad (2)$$

where g is another recurrent model, $s_0 = \mathbf{h_T}$, and y_0 is the start of sequence (SOS) token.

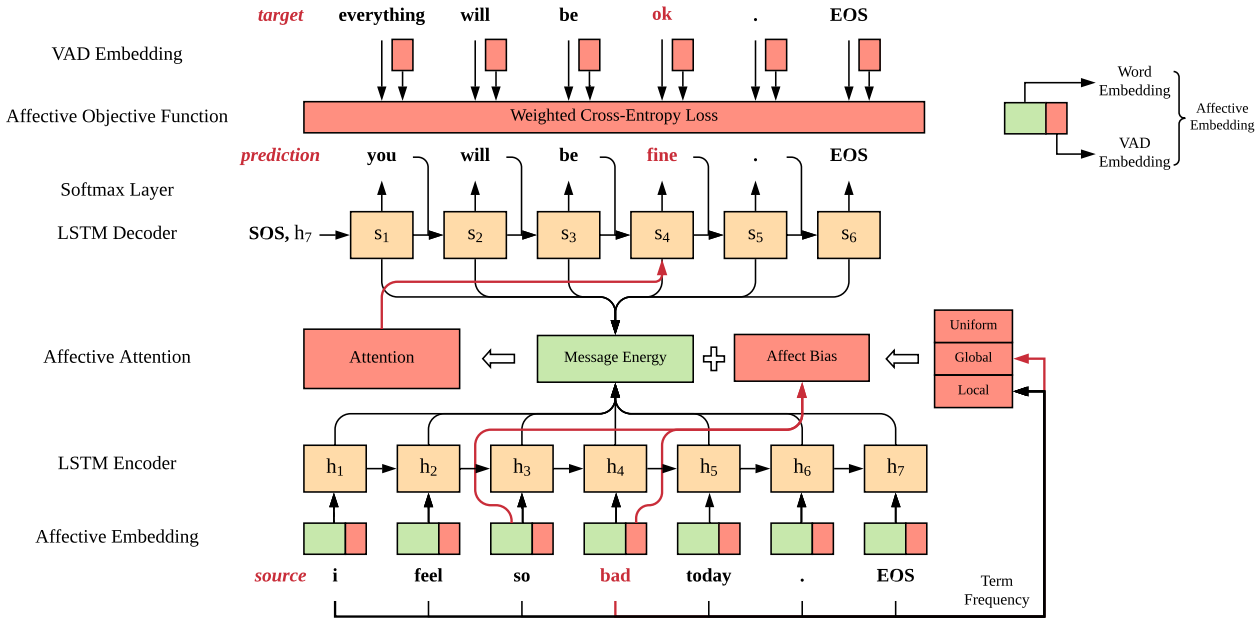


Figure 2: Overall architecture of our proposed AR-S2S. This diagram illustrates decoding "fine" and affect bias for "bad".

The output word probability in equation (1) is given by

$$p(y_{t'}) = softmax(\mathbf{W}^o \mathbf{s}_{t'}), \forall t^{'} = 1, 2, ..., T^{'}, \quad (3)$$

where $\mathbf{W}^o$ denotes a model parameter.

The attention mechanism (Luong, Pham, and Manning 2015) is proposed to solve the problem of limited representation power of the final input hidden state $\mathbf{h}_T$ on which the entire decoding process is conditioned. Specifically, the attention mechanism focuses on different parts of the input sequence by computing a context vector $\mathbf{c}_{t'}$ at each decoding time step $t^{'}, \forall t^{'} = 1, 2, ..., T^{'}$, as the weighted average of all input hidden states $\mathbf{h_t}, \forall t = 1, 2, ..., T$, as follows:

$$\mathbf{c}_{t'} = \sum_{t=1}^{T} \alpha_{t't} \mathbf{h}_t, \quad (4)$$

where the alignment vector $\alpha_{t't}$ is given by

$$\alpha_{t't} = \frac{\exp(e_{t't})}{\sum_{k=1}^{T} \exp(e_{t'k})}, \quad (5)$$

where $e_{t't} = score(\mathbf{h}_t, \mathbf{s}_{t'})$ is the message energy function that computes the energy or score between input hidden state $\mathbf{h}_t$ and output hidden state $\mathbf{s}_{t'}$. This message energy function is usually implemented as a Multilayer Perceptron (MLP). In our case, we use a simple dot product operation due to its fast training and good performance (Luong, Pham, and Manning 2015).

The context vector $\mathbf{c}_{t'}$ is then concatenated with the decoder hidden state $\mathbf{s}_{t'}$ to form an attentional hidden state $\hat{\mathbf{s}}_{\mathbf{t}'}$ as follows:

$$\hat{\mathbf{s}}_{\mathbf{t}'} = \tanh(\mathbf{W}^c[\mathbf{c}_{t'}; \mathbf{s}_{t'}]), \quad (6)$$

where $[;]$ denotes vector concatenation. Finally, $\hat{\mathbf{s}}_{\mathbf{t}'}$ replaces $\mathbf{s}_{t'}$ in equation (3) to compute the output word probability.

Dimensions	Values	Interpretations
Valence	3 - 7	pleasant - unpleasant
Arousal	3 - 7	low intensity - high intensity
Dominance	3 - 7	submissive - dominant

Table 1: Interpretations of clipped VAD embeddings.

Affect-Rich Seq2Seq Model

In this section, we present our proposed model to produce affect-rich responses, which falls outside the capability of vanilla Seq2Seq models. The overall model architecture is illustrated in Figure 2.

Affective Embedding

Our model adopts Valence, Arousal and Dominance (VAD) (Mehrabian 1996) embedding to encode word affects as vectors of size 3 from an annotated lemma-VAD pairs corpus (Warriner, Kuperman, and Brysbaert 2013). This corpus comprises 13,915 lemmas with VAD values annotated in the $[1, 9]$ scale. To leverage this corpus, we assign VAD values to words based on their lemmas. To increase coverage, we extend the corpus to 23,825 lemmas by assigning the average VAD values of their synonyms to absent lemmas. Furthermore, we empirically clip VAD values of all words to the $[3, 7]$ interval to prevent words with extreme VAD values from repeatedly showing in the generated responses, as observed in our preliminary experiments. The interpretations of clipped VAD embedding are presented in Table 1. For example, word "nice" is associated with the clipped VAD values: (V: 6.95, A: 3.53, D: 6.47). For words whose lemmas are not in the extended corpus, comprising approximately 10% of the entire training vocabulary, we assign them VAD values of $[5, 3, 5]$, which are the clipped values of a neutral

word. Note that a value of 3 in arousal (A) dimension is regarded as neutral because it has zero emotional intensity.

Finally, to remove bias, we normalize VAD embedding as $\overline{VAD}(x_t) = VAD(x_t) - [5, 3, 5]$, where $VAD(x_t) \in \mathbb{R}^3$ is the VAD embedding of word x_t. We incorporate VAD embedding by concatenation as follows:

$$\mathbf{e}(x_t) = [\mathbf{x_t}; \lambda \overline{VAD}(x_t)], \tag{7}$$

where $\mathbf{x_t} \in \mathbb{R}^m$ denotes the word embedding of x_t, $\mathbf{e}(x_t) \in \mathbb{R}^{m+3}$ denotes the final affective embedding of x_t, m denotes the dimensionality of word vectors, and $\lambda \in \mathbb{R}_+$ denotes the affect embedding strength hyper-parameter to tune the strength of VAD embeddings.

It is worth noting that the lemmas in our corpus were selected across multiple domains and are quite neutral (Brysbaert and New 2009). In addition, languages other than English, such as Spanish, Dutch, Finish, etc., also have such lemma-VAD pairs corpus, although in smaller sizes. Hence, our proposed conversational model has great potential to be directly applied to other languages.

Affective Attention

To incorporate affect into attention naturally, we make the intuitive assumption that humans pay extra attention on affect-rich words during conversations. Specifically, our model biases attention towards affect-rich words in the input sentences, as well as considers the effect of negators and intensifiers. Our model employs an affect bias η augmenting the affective strength of each word in the input sentences into the energy function (see equation (5)) as follows:

$$e_{t't} = \mathbf{h_t}^T \mathbf{s}_{t'} + \eta_t, \tag{8}$$

where $\mathbf{h_t}^T \mathbf{s}_{t'}$ denotes the conventional dot product energy function and η_t is defined as

$$\begin{aligned} \eta_t &= \gamma ||\mu(x_t)(1+\beta) \otimes \overline{VAD}(x_t)||_2^2, \\ \beta &= \tanh(\mathbf{W}^b \mathbf{x_{t-1}}), \end{aligned} \tag{9}$$

where $\otimes$ denotes element-wise multiplication, $||.||_k$ denotes l_k norm, $\mathbf{W}^b \in \mathbb{R}^{3 \times m}$ denotes a model parameter, $\beta \in \mathbb{R}^3$ is a scaling factor in V, A and D dimensions in the $[-1, 1]$ interval to scale the normalized VAD values of the current input word, $\gamma \in \mathbb{R}_+$ denotes the affective attention coefficient controlling the magnitude of affect bias towards affect-rich words in the input sentence, and $\mu(x_t) \in \mathbb{R}$ in the $[0, 1]$ interval denotes a measure of term importance of x_t (see the following paragraph).

Term Importance The introduction of term importance $\mu(x_t)$ as weights in computing affective attention is inspired by the sentence embedding work (Arora, Liang, and Ma 2016), where a simple weighted sum of word embedding algorithm with weights being smoothed inverse term frequency can achieve good performance in textual similarity tasks. Term frequency has been widely adopted in information retrieval to compute the importance of a word. In our model, we propose three approaches, namely "uniform importance" (ui), "global importance" (gi), and "local importance" (li) to compute $\mu(x_t)$:

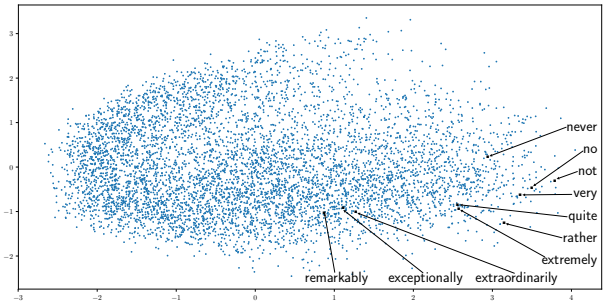


Figure 3: 2D plot of the most frequent 30,000 words in our training dataset in GloVe embedding after PCA. Selected common negators and intensifiers are annotated in text.

$$\mu(x_t) = \begin{cases} 1 & \text{ui} \\ a/(a + p(x_t)) & \text{gi} \\ \frac{\log(1/(p(x_t)+\epsilon))}{\sum_{t=1}^{T} \log(1/(p(x_t)+\epsilon))} & \text{li} \end{cases}, \tag{10}$$

where $p(x_t)$ denotes the term frequency of x_t in the training corpus, a denotes a smoothing constant that is usually set to 10^{-3} as suggested by Arora, Liang, and Ma (2016), and ϵ is another small smoothing constant with value 10^{-8}. We take the log function in $\mu_{li}(x_t)$ to prevent rare words from dominating the importance.

Modeling Negators and Intensifiers The introduction of β in equation (9) is to model the affect changes caused by negators and intensifiers. Often, negators make positive words negative but with much lower intensity, and make negative words less negative (Kiritchenko and Mohammad 2016). Thus, β is expected to be negative for negators because negators tend to reduce the affect intensity of the following word (e.g., "not bad"). Intensifiers usually adjust the intensities of positive words and negative words but do not flip their polarities (Carrillo-de Albornoz and Plaza 2013). As a result, β for extreme intensifiers (e.g., "extremely") is expected to be larger than β for less extreme intensifiers (e.g., "very"). To specifically consider these phenomena, β is modeled to be a nonlinear transformation through the word vector of x_{t-1}. This idea is inspired by the observation that common negators and intensifiers share some common underlying properties in their word vector representations. Figure 3 shows that several common negators and intensifiers tend to cluster together in 2D plots in GloVe embedding (Pennington, Socher, and Manning 2014) after applying Principle Component Analysis (PCA).

Note that our affective attention only considers unigram negators and intensifiers, however, they are empirically found as the majority of all negators and intensifiers. Statistics based on our training set indicate that the unigram intensifier "very" occurs 364,913 times, in comparison, the composite intensifier "not very" only occurs 2,838 times.

Affective Objective Function

The conventional objective function of seq2seq model is to maximize the probability of target response Y given input

sequence X measured by cross-entropy loss. To encourage the generation of affect-rich words, we incorporate VAD embedding of words into cross-entropy loss as follows:

$$\Psi_{t'} = -|V| \frac{1 + \delta||\overline{VAD}(y_{t'})||_2}{\sum_{\hat{y}_{t'} \in V}(1 + \delta||\overline{VAD}(\hat{y}_{t'})||_2)} \log(p(y_{t'})), \tag{11}$$

where $t^{'} = 1, 2, ..., T^{'}$, $\Psi_{t'}$ denotes the affective loss at decoding time step $t^{'}$, $y_{t'}$ denotes the target token at decoding time step $t^{'}$, V denotes the dataset vocabulary, and δ denotes a hyper-parameter named affective loss coefficient, which regulates the contribution of VAD embedding.

Our proposed affective loss is essentially a weighted cross-entropy loss. The weights are constant and positively correlated with VAD strengths in l_2 norm. The weight normalization is applied to ensure that our weights do not alter the overall learning rate during optimization. Intuitively, our affective loss encourages affect-rich words to obtain higher output probability, which effectively introduces a probability bias into the decoder language model towards affect-rich words. This bias is controlled by our affective loss coefficient δ. When $\delta = 0$, our affective loss falls back to the conventional unweighted cross-entropy loss.

It is worth noting that our weighted cross-entropy loss incorporating external word knowledge, i.e., VAD in our case, is simple but effective in controlling the response style. Our loss function has many other potential application areas such as controlled neural text generation.

Experimental Evaluation

In this section, we present our datasets, evaluation methods, experimental results and discussions. Following the experimental setup presented in (Zhou et al. 2018), we conduct **model component test (MCT)** to examine the effectiveness of our proposed affective attention and affective objective function in generating affect-rich responses. In addition, we conduct **preference test (PT)** between our best model (**AR-S2S**) and the state-of-the-art baseline of comparable model size to compare model responses. Finally, we conduct **sensitivity analysis** on the hyper-parameters introduced in our model to analyze their impacts on language fluency and the number of distinct affect-rich words produced.

Datasets

We use OpenSubtitles dataset (Tiedemann 2009) as our training dataset due to its large size. We use relatively less noisy Cornell Movie Dialog Corpus dataset (Danescu-Niculescu-Mizil and Lee 2011) as our validation dataset for more reliable tuning. We use DailyDialog dataset (Li et al. 2017) for testing to examine model generalizations in different corpus domains.

The pairs in the training dataset are selected by a simple rule that the input sentence ends with a question mark and the time interval between the pair of input and output sentences is less than 20 seconds. In addition, sound sequences such as "*BANG*" are removed. These pairs are then expanded (e.g., isn't $\rightarrow$ is not), tokenized, and special symbols and numbers were removed. Finally, the pairs with either input or output sentence longer than 20 words are removed. The validation and testing datasets are preprocessed by word expansion, tokenization and removal of special symbols and numbers. Since we are modeling single-turn dialogue system, only the first two utterances from each dialogue session in the testing dataset are extracted because using utterances in the middle would require context to respond.

After data preprocessing, we randomly select 5 million pairs from OpenSubtitles dataset as the training dataset with a vocabulary comprising the most 30,000 frequent words, covering 98.89% of all tokens. We randomly sample 100K pairs from Cornell Movie Dialog Corpus dataset for validation and 10K pairs from DailyDialog dataset for testing.

Evaluation Methods

We adopt perplexity metric to measure the language fluency of a conversational model, as it is the only well-established automatic evaluation method in conversation modeling. Other metrics such as BLEU (Papineni et al. 2002) do not correlate well with human judgments (Liu et al. 2016). A model with lower perplexity indicates that it is more confident about the generated responses. Note that a model with low perplexity does not guarantee to be a good conversational model because it may achieve so by always generating short responses.

To qualitatively examine model performance, we conduct widely adopted human evaluations. We randomly sample 100 input sentences from the testing dataset. For each input sentence, we then randomize the order of the responses generated by each comparison model. For each response, five human annotators are asked to evaluate two aspects:

- **+2**: (content) The response has correct grammar and is relevant and natural / (emotion) The response has adequate and appropriate emotions conveyed.
- **+1**: (content) The response has correct grammar but is too universal / (emotion) The response has inadequate but appropriate emotions conveyed.
- **0**: (content) The response has either grammar errors or is completely irrelevant / (emotion) The response has either no or inappropriate emotions conveyed.

Experiment 1: Model Component Test (MCT)

We compare the following models to examine the performance of our proposed affective attention and affective objective function on model perplexity and human evaluations:

S2S: The standard Seq2Seq model with attention.

S2S-UI, **S2S-GI**, **S2S-LI**: The standard Seq2Seq model with our proposed affective attention using μ_{ui}, μ_{gi} and μ_{li} (see equation (10)), respectively.

S2S-AO: The standard Seq2Seq model with attention and our proposed affective objective function (see equation (11)).

AR-S2S: our best model, which incorporates both μ_{li} and affective objective function.

All models have a word embedding of size 1027 (1024 + 3) and hidden size of 1024. Both encoder and decoder have two layers of bi-directional LSTM. All models implement

Experiment	Model	#Params	PPL†	PPL‡
MCT (5M pairs)	S2S	99M	42.5	124.3
	S2S-UI	99M	40.4	116.4
	S2S-GI	99M	40.7	120.3
	S2S-LI	99M	40.4	117.0
	S2S-AO	99M	40.2	115.7
	AR-S2S	99M	**39.8**	**113.7**
PT (3M pairs)	S2S	66M	41.2	130.6
	S2S-Asghar	66M	46.4	137.2
	AR-S2S	66M	**40.3**	**121.0**

Table 2: Model test perplexity. Symbol † indicates in-domain perplexity obtained on 10K test pairs from the OpenSubtitles dataset. Symbol ‡ indicates out-domain perplexity obtained on 10K test pairs from the DailyDialog dataset.

affective embedding. Parameters λ, δ and a are set to 0.1, 0.15 and 10^{-3}, respectively. Parameter γ for S2S-UI, S2S-GI and S2S-LI are set to 0.5, 1 and 5, respectively. The beam size is set to 20. Note that all models implement the maximum mutual information (MMI) objective function (Li et al. 2016) during inference to levitate the problem of generic responses (e.g., "I don't know"). For all models, a simple re-rank operation is applied during inference to rank the generated responses $\hat{Y}$ based on their affective strength computed as $\frac{1}{|\hat{Y}|}\sum_{y\in\hat{Y}}||\overline{VAD}(y)||_2$. All models are initialized with a uniform distribution in the $[-0.08, 0.08]$ interval, using the same seed. We trained all models with a batch size of 64 for 5 epochs using Adam (Kingma and Ba 2014) optimization ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) with the learning rate of 0.0001 throughout the training process.

Results Table 2 presents the results on model test perplexity in both MCT and PT (see Experiment 2). To analyze model generalization in different domains, we additionally report test perplexity on in-domain test dataset, which is created using 10K test pairs from the OpenSubtitles dataset. All models have comparable perplexity on both in-domain and out-domain test datasets, empirically showing that our proposed methods do not cause performance degradation in language fluency. One note is that the out-domain test perplexity for all models is quite large as compared to in-domain perplexity, as well as other dialog systems, e.g., (Vinyals and Le 2015). One possible reason is that our testing dataset is different from the training dataset in terms of both vocabulary and linguistic distributions (the former was created from daily conversations, whereas the latter was created from movie subtitles). As a result, the models may not generalize well.

Tables 3 and 4 present the evaluation results in MCT by five human annotators on the content quality and emotion quality, respectively. The values in brackets denote performance improvement in percentage. The Fleiss' kappa (Fleiss and Cohen 1973) for measuring inter-rater agreement is included as well. All models have "moderate agreement" or "substantial agreement". For content quality, all models except S2S-AO have noticeably more +2 ratings than S2S. For emotion quality, it is clear that all of our proposed affective models have significant improvement over S2S. Among

Model (%)	+2	+1	0	Score	Kappa
S2S	22.4	47.0	30.6	0.918	0.544
S2S-UI	**30.0**	48.6	21.4	**1.086 (+18.3%)**	0.458
S2S-GI	28.6	46.6	24.8	1.038 (+13.1%)	0.413
S2S-LI	29.4	47.2	23.4	1.060 (+15.5%)	0.525
S2S-AO	25.0	46.0	29.0	0.960 (+4.3%)	0.482
AR-S2S	29.6	44.8	25.6	1.040 (+13.3%)	0.487

Table 3: Human evaluations on content quality (MCT).

Model (%)	+2	+1	0	Score	Kappa
S2S	19.0	33.2	47.8	0.712	0.613
S2S-UI	23.6	36.0	40.4	0.832 (+16.9%)	0.483
S2S-GI	26.0	34.2	39.8	0.862 (+21.1%)	0.652
S2S-LI	24.6	36.4	39.0	0.856 (+20.2%)	0.706
S2S-AO	22.6	37.6	39.8	0.828 (+16.3%)	0.602
AR-S2S	**26.8**	37.2	36.0	**0.908 (+27.5%)**	0.625

Table 4: Human evaluations on emotion quality (MCT).

Message	Model	Response
Mommy, can i stay up until eleven of the clock?	S2S (MCT)	Of course you can stay up late.
	AR-S2S (MCT)	Of course you can, **sweetheart**.
You are home late today, david. How was school?	S2S (MCT)	It was fine.
	AR-S2S (MCT)	**Great fun** today.
Do you like singing?	S2S (PT)	Yes, i do.
	S2S-Asghar (PT)	I do not know.
	AR-S2S (PT)	I **love** music.
I'm pretty sure that jim will turn out to be a good lawyer.	S2S (PT)	He will turn out to be a good lawyer.
	S2S-Asghar (PT)	I'm sure he will.
	AR-S2S (PT)	The **best** lawyer in the world.

Table 5: Sample responses for models in both MCT and PT. Text in bold are affect-rich words.

the three affective attention mechanisms, S2S-LI achieves the best overall performance. Note that the improvement gained by affective attention and affective objective function are partially orthogonal. One explanation is that by actively paying attention to affect-rich words in the input sentence, our model is able to produce more accurate affect-rich words during decoding. Therefore, combing both techniques (AR-S2S) results in maximum improvement in emotion quality. Table 5 presents some sample responses in the testing dataset.

Analysis of Affective Attention To examine our hypothesis that our affective attention mechanism can correctly capture the effect of negators and intensifiers, we plot the learned parameter β (see equation (9)) in the Valence and Arousal dimensions in Figure 4. It is obvious that our model successfully learned to make β negative for negators. In addition, several extreme intensifiers such as "exceptionally" and "remarkably" have higher β than less extreme intensi-

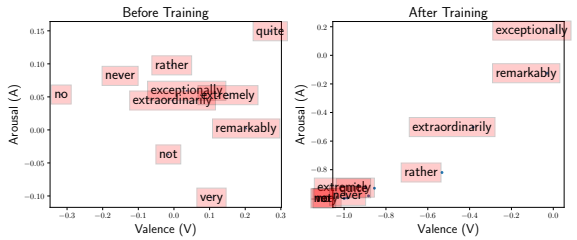


Figure 4: Learned parameter β (see equation (9)) in Valence (V) and Arousal (A) dimensions for several common negators and intensifiers. Left sub-figure: before AR-S2S is trained. Right sub-figure: after AR-S2S is trained.

Figure 5: Learned attention on a sample input sentence from the testing dataset. From top to bottom, the models are S2S, S2S-UI, S2S-GI and S2S-LI, respectively. Darker colors indicate larger strength.

fiers such as "very" and "quite", which is consistent with our hypothesis. One note is that our model does not learn well for some intensifiers such as "extremely", whose β is comparable to less extreme intensifiers such as "very". This result is not surprising because the impacts of intensifiers are difficult to be completely captured as they tend to vary depending on the following words (Kiritchenko and Mohammad 2016).

Figure 5 shows the attention strength over a sample input sentence in the testing dataset. As expected, our proposed affective attention models place extra attention on affect-rich words, i.e., "good" in this case. In addition, S2S-UI and S2S-LI have larger strengths than S2S-GI. This result is aligned with our model's assumption because different "term importance" have different impacts on the attention strengths and the word "good" here is quite common ($p(\text{"}good\text{"}) = 0.00143$), which leads to the lower strength in S2S-GI.

Analysis of Affective Objective Function We analyze the capability of our proposed affective objective function in producing affect-rich words. Table 6 presents the number of distinct affect-rich words in randomly selected 1K test responses produced by different models. Affect-rich words are defined as words with VAD strength in l_2 norm exceeding the given threshold. It is clear that all S2S-AO models can produce more affect-rich words than S2S. In addition, the number of affect-rich words for every threshold increases steadily as the affective objective coefficient δ increases, showing a good controllability of our model via δ.

	Threshold for l_2 Norm of VAD		
Model	3	2	1
S2S	25	104	190
S2S-AO ($\delta = 0.5$)	36	138	219
S2S-AO ($\delta = 1$)	50	154	234
S2S-AO ($\delta = 2$)	69	177	256

Table 6: Number of distinct affect-rich words (MCT).

	Threshold for l_2 Norm of VAD		
Model	3	2	1
S2S	21	83	157
S2S-Asghar	31	120	217
AR-S2S	52	173	319

Table 7: Number of distinct affect-rich words (PT).

Experiment 2: Preference Test (PT)

We conduct human preference test to compare our **AR-S2S** with the state-of-the-art baseline **S2S-Asghar**, the best model proposed in (Asghar et al. 2018). To the best of our knowledge, S2S-Asghar is the only model in the neural conversational model literature that aims to produce affect-rich responses in an end-to-end manner (i.e., without explicit user-input emotions). We also include **S2S** for comparison.

To make comparisons fair, we follow the specifications of S2S-Asghar and keep the number of parameters in all models comparable by reducing the size of our model. We use a smaller training dataset with 3 million random pairs and a vocabulary of size 20,000 due to the reduced model size. Note that our training dataset is still significantly larger than the original dataset used in (Asghar et al. 2018), which comprises only 300K pairs and a vocabulary size of 12,000. All models have a word embedding of size 1027, a single-layer LSTM encoder and a single-layer LSTM decoder. All training specifications remain the same as the MCT except that S2S-Asghar is trained for 4 epochs with conventional cross-entropy loss and 1 more epoch with their proposed objective function, which includes a term to maximize affective content.

For human evaluation, we follow the same procedures as adopted in MCT except that five human annotators were asked to choose their preferred responses based on content quality and emotion quality, respectively, instead of annotating +2, +1 and 0. Ties are allowed.

Results Table 7 shows the number distinct of affect-rich words in randomly selected 1K responses produced by S2S, S2S-Asghar and our model. It is clear that our model produces significantly more affect-rich words than both S2S-Asghar and S2S.

Table 8 shows the result of human evaluation. The Fleiss' kappa scores for content/emotion qualities are included in the last column. All models have "moderate agreement" or "substantial agreement". For content preference, our model scores relatively 21% higher than S2S-Asghar. For emotion preference, our model scores relatively 50% higher than S2S-Asghar. These findings show that our model is capable of producing better responses that are not only more ap-

Model (%)	Content	Emotion	Kappa
S2S	64	26	0.522/0.749
S2S-Asghar	66 (+3.1%)	32 (+23.1%)	0.554/0.612
AR-S2S	**80 (+25.0%)**	**49 (+88.5%)**	0.619/0.704

Table 8: Human preference test (PT).

propriate in syntax and content, but also significantly more affect-rich than the state-of-the-art model.

Experiment 3: Sensitivity Analysis

We examine the impacts of the affect embedding strength λ, affective attention hyper-parameter γ, as well as affective loss hyper-parameter δ on model perplexity and the number of affect-rich words produced. Due to the large number of experiments, we conduct the sensitivity analysis using 1 million pairs and a vocabulary of size 20,000. All training specifications remain the same as MCT except that the number of LSTM layers is 1, the hidden layer size is 512 and the embedding layer size is 303.

Results Figure 6 shows the plots of model test perplexity versus λ, γ and δ. Our model is fairly robust to a wide range of λ, γ and δ, regardless of the type of term importance. It is worth noting that the generated responses tend to become shorter with $\gamma \in [20, \infty]$, which may be caused by excessive attention placed on affect-rich words during decoding. Another interesting finding is that our affective objective function slightly improves test perplexity. One possible explanation is that affect-rich words are less common than generic words in our training corpus. As a result, our weighted cross-entropy loss placing extra weights on them improves the overall prediction performance.

Figure 7 shows the plots of the number of distinct affect-rich words in randomly selected 1K test responses versus λ, γ and δ. The number of distinct words increases slightly when λ increases from 0 to 0.3, and then gradually decreases and stabilizes as λ increases from 0.3 to 1. For γ in all three term importance, there is an initial boost in the number of distinct words when γ is small, i.e., $\gamma \in [0, 5]$. However, as γ furher increases, the number of distinct words gradually decreases, which may be caused by limited word space during decoding due to excessive attention on affect-rich words. Among the three term importance proposed, local importance (μ_{li}) is slightly more robust against γ than the other two approaches. Finally, the number of distinct words consistently increases with δ, which is similar to our findings from Table 6. Note that the numbers in this sensitivity analysis are much smaller than MCT, which can be attributed to smaller models and less training examples.

Conclusion

In this paper, we propose an end-to-end open-domain neural conversational model that produces affect-rich responses without performance degradation in language fluency. Our model leverages external word-VAD knowledge to encode affect information into the conversational model. In addition, our model captures user emotions by paying extra attention to affect-rich words in input sentences and considering the effect caused by negators and intensifiers. Lastly, our model is trained with an affect-incorporated weighted cross-entropy loss to encourage the generation of affect-rich words. Empirical studies on both model perplexity and human evaluations show that our model outperforms the state-of-the-art model of comparable size in producing natural and affect-rich responses.

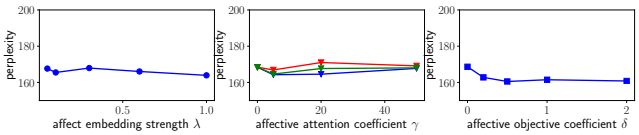


Figure 6: Sensitivity analysis for affect embedding strength λ, affective attention coefficient γ, and affective objective coefficient δ on model perplexity. The blue, red and green curves (*best viewed in color*) in the middle sub-figure represent μ_{ui}, μ_{gi} and μ_{li} (see equation (10)), respectively.

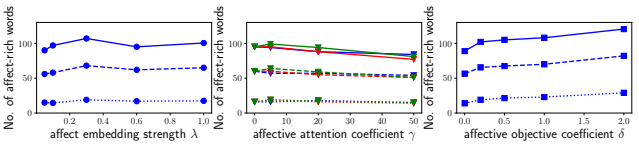


Figure 7: Sensitivity analysis for affect embedding strength λ, affective attention coefficient γ, and affective objective coefficient δ on the number of distinct affect-rich words in randomly selected 1K test responses. The solid, dashed and dotted curves correspond to l_2 norm threshold of 1, 2 and 3, respectively. The blue, red and green curves (*best viewed in color*) in the middle sub-figure represent μ_{ui}, μ_{gi} and μ_{li} (see equation (10)), respectively.

Acknowledgments

This research is supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative and 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

[Arora, Liang, and Ma 2016] Arora, S.; Liang, Y.; and Ma, T. 2016. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*.

[Asghar et al. 2018] Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; and Mou, L. 2018. Affective neural response generation. In *ECIR*, 154–166.

[Brysbaert and New 2009] Brysbaert, M., and New, B. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods* 41(4):977–990.

[Callejas, Griol, and López-Cózar 2011] Callejas, Z.; Griol, D.; and López-Cózar, R. 2011. Predicting user mental states in

spoken dialogue systems. *EURASIP Journal on Advances in Signal Processing* 2011(1):6.

[Carrillo-de Albornoz and Plaza 2013] Carrillo-de Albornoz, J., and Plaza, L. 2013. An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *JASIST* 64(8):1618–1633.

[Cho et al. 2014] Cho, K.; van Merriënboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, 1724–1734.

[Danescu-Niculescu-Mizil and Lee 2011] Danescu-Niculescu-Mizil, C., and Lee, L. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

[Fitzpatrick, Darcy, and Vierhile 2017] Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health* 4(2):e19.

[Fleiss and Cohen 1973] 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 et al. 2018] Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; tau Yih, W.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

[Ghosh et al. 2017] 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.

[Hewstone and Stroebe 2001] Hewstone, M., and Stroebe, W. 2001. *Introduction to Social Psychology: A European Perspective*. Oxford Blackwell Publishers.

[Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

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

[Kiritchenko and Mohammad 2016] Kiritchenko, S., and Mohammad, S. 2016. The effect of negators, modals, and degree adverbs on sentiment composition. In *WASSA*, 43–52.

[Li et al. 2016] Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*, 110–119.

[Li et al. 2017] Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, 986–995.

[Liu et al. 2016] Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2122–2132.

[Luong, Pham, and Manning 2015] Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, 1412–1421.

[Martinovski and Traum 2003] Martinovski, B., and Traum, D. 2003. Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Dialogue Systems*, 11–16.

[Mehrabian 1996] Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14(4):261–292.

[Ochs, Pelachaud, and Sadek 2008] Ochs, M.; Pelachaud, C.; and Sadek, D. 2008. An empathic virtual dialog agent to improve human-machine interaction. In *AAMAS*, 89–96.

[Papineni et al. 2002] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

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

[Picard 1997] Picard, R. W. 1997. *Affective computing*, volume 252. MIT Press Cambridge.

[Robison, McQuiggan, and Lester 2009] Robison, J.; McQuiggan, S.; and Lester, J. 2009. Evaluating the consequences of affective feedback in intelligent tutoring systems. In *ACII*, 1–6.

[Serban et al. 2017] Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 3295–3301.

[Shang, Lu, and Li 2015] Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *ACL*, 1577–1586.

[Skowron 2010] Skowron, M. 2010. Affect listeners: Acquisition of affective states by means of conversational systems. In *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer. 169–181.

[Sutskever, Vinyals, and Le 2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*, 3104–3112.

[Tang et al. 2017] Tang, C.; Wang, D.; Tan, A.-H.; and Miao, C. 2017. EEG-based emotion recognition via fast and robust feature smoothing. In *BI*, 83–92.

[Tiedemann 2009] Tiedemann, J. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, 237–248.

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

[Wang, Tan, and Miao 2016] Wang, D.; Tan, A.-H.; and Miao, C. 2016. Modeling autobiographical memory in human-like autonomous agents. In *AAMAS*, 845–853.

[Warriner, Kuperman, and Brysbaert 2013] Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods* 45(4):1191–1207.

[Xing et al. 2017] 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.

[Zhao, Zhao, and Eskenazi 2017] Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, 654–664.

[Zhou and Wang 2018] Zhou, X., and Wang, W. Y. 2018. Mojitalk: Generating emotional responses at scale. In *ACL*, 1128–1137.

[Zhou et al. 2018] Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.