Generating Diversified Comments via Reader-Aware Topic Modeling and Saliency Detection

Wei Wang[1,2,*] **, Piji Li**[3]**, Hai-Tao Zheng**[1,2,†]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Department of Computer Science and Technology, Tsinghua University
[3]Tencent AI Lab
w-w16@mails.tsinghua.edu.cn, pijili@tencent.com
zheng.haitao@sz.tsinghua.edu.cn

Abstract

Automatic comment generation is a special and challenging task to verify the model ability on news content comprehension and language generation. Comments not only convey salient and interesting information in news articles, but also imply various and different reader characteristics which we treat as the essential clues for **diversity**. However, most of the comment generation approaches only focus on saliency information extraction, while the reader-aware factors implied by comments are neglected. To address this issue, we propose a unified reader-aware topic modeling and saliency information detection framework to enhance the quality of generated comments. For reader-aware topic modeling, we design a variational generative clustering algorithm for latent semantic learning and topic mining from reader comments. For saliency information detection, we introduce Bernoulli distribution estimating on news content to select saliency information. The obtained topic representations as well as the selected saliency information are incorporated into the decoder to generate diversified and informative comments. Experimental results on three datasets show that our framework outperforms existing baseline methods in terms of both automatic metrics and human evaluation. The potential ethical issues are also discussed in detail.

Introduction

For natural language generation research, automatic comment generation is a challenging task to verify the model ability on aspects of news information comprehension and high-quality comment generation (Reiter and Dale 1997; Gatt and Krahmer 2018; Qin et al. 2018; Huang et al. 2020).

One common phenomenon is that, for the same news article, there are usually hundreds or even thousands of different comments proposed by readers in different backgrounds. Figure 1 depicts an example of a news article (truncated) and three corresponding comments from Yahoo. The news article is about "The Walking Dead". From the example, we can observe and conclude two characteristics: (1) Although the news article depicts many different aspects of the event and conveys lots of detailed information, readers usually pay attention to part of the content, which means that not all the content information is **salient** and **important**. As shown in Figure 1, the first two readers both focus on "The Walking Dead" and the third reader focuses on "Reedus". None of them mention other details in the content. (2) Different readers are usually interested in different topics, and even on the same topic, they may hold different opinions, which makes the comments **diverse** and **informative**. For example, the first reader gives the comment from the topic of "feeling" and he "cannot accept Lincoln's leaving". The second reader comments on the topic of "plot" and thinks "it is old-fashioned". The third reader comments from the topic of "acting", saying that "Reedus can't play that role".

Title: Andrew Lincoln Poised To Walk Off 'The Walking Dead' Next Season; Norman Reedus To Stay

Body (truncated): He has lost his on-screen son, his wife and a number of friends to the zombie apocalypse, and now The Walking Dead star Andrew Lincoln looks to be taking his own leave of the AMC blockbuster. Cast moves on the series will also see Norman Reedus stay put under a new $ 20 million contract.
Almost unbelievable on a show where almost no one is said to be safe, the man who has played Rick Grimes could be gone by the end of the upcoming ninth season... The show will reset with Reedus ' Daryl Dixon even more in the spotlight ...

Comment A: I'm not watching TWD without Lincoln.
Comment B: Storylines getting stale and they keep having the same type trouble every year.
Comment C: Reedus ca n't carry the show playing Daryl.

Figure 1: A news example from Yahoo.

Therefore, news comments are produced based on the interactions between readers and news articles. Comments not only imply different important information in news articles, but also convey distinct reader characteristics. Precisely because of these reader-aware factors, we can obtain a variety of diversified comments under the same news article. Intuitively, as the essential reason for diversity, these reader-aware factors should be considered jointly with saliency information detection in the task of diversified comment generation. However, it is rare that works consider these two components simultaneously. Zheng et al.(2017) proposed to

[*]Work was done during internship at Tencent AI Lab.
[†]Corresponding author.

generate one comment only based on the news title. Qin et al.(2018) extended the work to generate a comment jointly considering the news title and body content. These two seq2seq (Sutskever, Vinyals, and Le 2014) based methods conducted saliency detection directly via attention modeling. Li et al.(2019) extracted keywords as saliency information. Yang et al.(2019) proposed a reading network to select important spans from the news article. Huang et al.(2020) employed the LDA (Blei, Ng, and Jordan 2003) topic model to conduct information mining from the content. All these works concern the saliency information extraction and enhance the quality of comments. However, various reader-aware factors implied in the comments, which are the essential factors for diversity as well, are neglected.

To tackle the pre-mentioned issues, we propose a reader-aware topic modeling and saliency detection framework to enhance the quality of generated comments. The goal of reader-aware topic modeling is to conduct reader-aware latent factors mining from the **comments**. The latent factors might be either reader interested topics or the writing styles of comments, or more other detailed factors. We do not design a strategy to disentangle them, instead, we design a unified latent variable modeling component to capture them. Specifically, inspired by Jiang et al. (2017), we design a Variational Generative Clustering (VGC) model to conduct latent factors modeling from the reader comments. The obtained latent factor representations can be interpreted as news topics, user interests, or writing styles. For convenience, we collectively name them as **Topic**. For reader-aware saliency information detection, we build a saliency detection component to conduct the Bernoulli distribution estimating on the news content. Gumbel-Softmax is introduced to address the non-differentiable sampling operation. Finally, the obtained topic representation vectors and the selected saliency news content are integrated into the generation framework to control the model to generate diversified and informative comments.

We conduct extensive experiments on three datasets in different languages: NetEase News (Chinese) (Zheng et al. 2017), Tencent News (Chinese) (Qin et al. 2018), and Yahoo! News (English) (Yang et al. 2019). Experimental results demonstrate that our model can obtain better performance according to automatic evaluation and human evaluation. In summary, our contributions are as follows:

- We propose a framework to generate diversified comments jointly considering saliency news information detection and reader-aware latent topic modeling.
- We design a Variational Generative Clustering (VGC) based component to learn the reader-aware latent topic representations from comments.
- For reader-aware saliency information detection, Bernoulli distribution estimating is conducted on the news content. Gumbel-Softmax is introduced to address the non-differentiable sampling operation.
- Experiments on three datasets demonstrate that our model outperforms state-of-the-art baseline methods according to automatic evaluation and human evaluation.

Methodology

Overview

To begin with, we state the problem of news comment generation as follows: given a news title $T = \{t_1, t_2, \ldots, t_m\}$ and a news body $B = \{b_1, b_2, \ldots, b_n\}$, the model needs to generate a comment $Y = \{y_1, y_2, \ldots, y_l\}$ by maximizing the conditional probability $p(Y|X)$, where $X = [T, B]$. As shown in Figure 2, the backbone of our work is a sequence-to-sequence framework with attention mechanism (Bahdanau, Cho, and Bengio 2014). Two new components of reader-aware topic modeling and saliency information detection are designed and incorporated for better comment generation. For reader-aware topic modeling, we design a variational generative clustering algorithm for latent semantic learning and topic mining from reader comments. For reader-aware saliency information detection, we introduce a saliency detection component to conduct the Bernoulli distribution estimating on news content. Finally, the obtained topic vectors as well as the selected saliency information are incorporated into the decoder to conduct comment generation.

The Backbone Framework

The backbone of our work is a sequence-to-sequence framework with attention mechanism (Bahdanau, Cho, and Bengio 2014). A BiLSTM encoder is used to encode the input content words X into vectors. Then a LSTM-based decoder generates a comment Y conditioning on a weighted content vector computed by attention mechanism. Precisely, a word embedding matrix is used to convert content words into embedding vectors. Then these embedding vectors are fed into encoder to compute the forward hidden vectors via:

$$\overrightarrow{\mathbf{h}_i} = LSTM_f^e\left(\mathbf{x}_i, \overrightarrow{\mathbf{h}}_{i-1}\right), \tag{1}$$

where $\overrightarrow{\mathbf{h}_i}$ is a d-dimension hidden vector and $LSTM_f^e$ denotes the LSTM unit. The reversed sequence is fed into $LSTM_b^e$ to get the backward hidden vectors. We concatenate them to get the final hidden representations of content words. And the representation of news is : $\mathbf{h}^e = \left[\overrightarrow{\mathbf{h}_{|X|}}; \overleftarrow{\mathbf{h}_0}\right]$.

The state of decoder is initialized with $\mathbf{h}^e$. For predicting comment word y_t, the hidden state is first obtained by:

$$\mathbf{h}_t = LSTM_d\left(\mathbf{y}_{t-1}, \mathbf{h}_{t-1}\right), \tag{2}$$

where $h_t \in \mathbb{R}^{2d}$ is the hidden vector of comment word and $LSTM_d$ is the LSTM unit. Then we use attention mechanism to query the content information from source input. The weight of each content word is computed as follows:

$$\begin{aligned} e_{ti} &= \mathbf{h}_t^\top W_a \mathbf{h}_i^e, \\ \alpha_{ti} &= \frac{\exp\left(e_{ti}\right)}{\sum_{i\prime=1}^{|X|} \exp\left(e_{ti\prime}\right)}, \end{aligned} \tag{3}$$

where $W_a \in \mathbb{R}^{2d \times 2d}$, $\mathbf{h}_i^e$ is the hidden vector of content word i, α_{ti} is the normalized attention score on x_i at time step t. Then the attention-based content vector is obtained by: $\tilde{\mathbf{h}}_t^e = \sum_i \alpha_{ti} \mathbf{h}_i^e$. The hidden state vector is updated by:

$$\tilde{\mathbf{h}}_t = \tanh(W_c[\mathbf{h}_t; \tilde{\mathbf{h}}_t^e]). \tag{4}$$

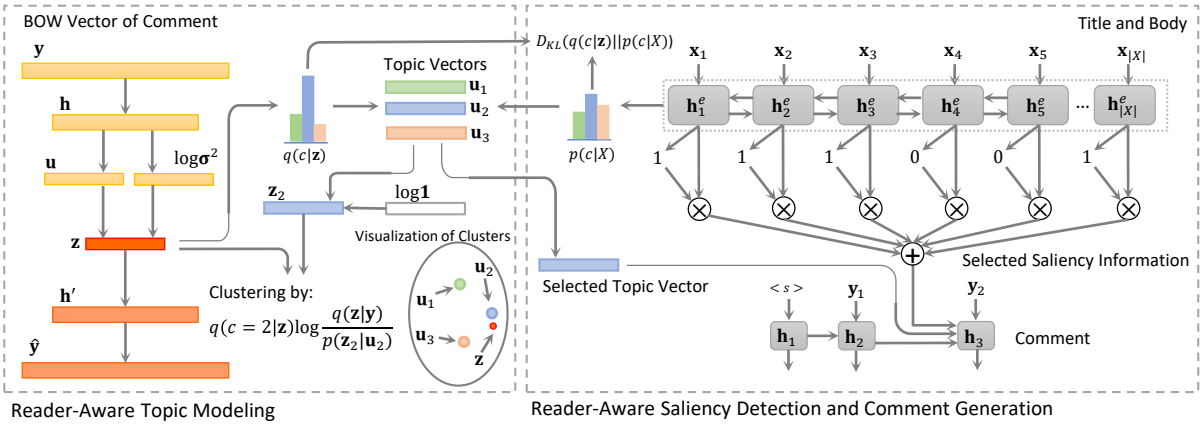

Figure 2: The framework of our proposed method. Left: the reader-aware topic modeling component is used to learn topic vectors. Specifically, the comment $\mathbf{y}$ is encoded to get latent semantic vector $\mathbf{z}$. Then the classifier $q(c|\mathbf{z})$ classifies $\mathbf{z}$ into one topic. In addition, $\mathbf{z}$ is decoded to reconstruct the comment. Topic vectors are learned according to Equation 9. Right: the reader-aware saliency detection is used to select saliency words and the topic selector $p(c|X)$ is used to select an appreciate topic vector. Finally the comment is generated conditioning on the selected topic vector and selected saliency information.

Finally, the probability of the next word y_t is computed via:

$$p\left(y_t|y_{t-1}, X\right) = \text{softmax}\left(linear\left(\tilde{\mathbf{h}}_t\right)\right), \tag{5}$$

where $linear(\cdot)$ is a linear transformation function.

During training, cross-entropy loss $\mathcal{L}_{ce}$ is employed as the optimization objective.

Reader-Aware Topic Modeling

Reader-aware topic modeling is conducted on all the comment sentences, aiming to learn the reader-aware topic representations only from the comments. To achieve this goal, we design a variational generative clustering algorithm which can be trained jointly with the whole framework in an end-to-end manner.

Since this component is a generative model, thus for each comment sentence Y (we employ Bag-of-Words feature vector $\mathbf{y}$ to represent each comment sentence), the generation process is: (1) A topic c is generated from the prior topic categorical distribution $p(c)$; (2) A latent semantic vector $\mathbf{z}$ is generated conditionally from a Gaussian distribution $p(\mathbf{z}|c)$; (3) $\mathbf{y}$ is generated from the conditional distribution $p(\mathbf{y}|\mathbf{z})$. According to the generative process above, the joint probability $p(\mathbf{y}, \mathbf{z}, c)$ can be factorized as:

$$p(\mathbf{y}, \mathbf{z}, c) = p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|c)p(c). \tag{6}$$

By using Jensen's inequality, the log-likelihood can be written as:

$$\begin{aligned} \log p(\mathbf{y}) &= \log \int_{\mathbf{z}} \sum_{c} p(\mathbf{y}, \mathbf{z}, c) d\mathbf{z} \\ &\geq E_{q(\mathbf{z},c|\mathbf{y})}\left[\log \frac{p(\mathbf{y}, \mathbf{z}, c)}{q(\mathbf{z}, c|\mathbf{y})}\right] = \mathcal{L}_{\text{ELBO}}(\mathbf{y}), \end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{ELBO}}$ is the evidence lower bound (ELBO), $q(\mathbf{z}, c|\mathbf{y})$ is the variational posterior to approximate the true posterior $p(\mathbf{z}, c|\mathbf{y})$ and can be factorized as follows:

$$q(\mathbf{z}, c|\mathbf{y}) = q(\mathbf{z}|\mathbf{y})q(c|\mathbf{z}). \tag{8}$$

Then based on Equation 6 and Equation 8, the $\mathcal{L}_{\text{ELBO}}$ can be rewritten as:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\mathbf{y}) =& E_{q(\mathbf{z},c|\mathbf{y})}\left[\log \frac{p(\mathbf{y}, \mathbf{z}, c)}{q(\mathbf{z}, c|\mathbf{y})}\right] \\ =& E_{q(\mathbf{z},c|\mathbf{y})}[\log p(\mathbf{y}, \mathbf{z}, c) - \log q(\mathbf{z}, c|\mathbf{y})] \\ =& E_{q(\mathbf{z},c|\mathbf{y})}[\log p(\mathbf{y}|\mathbf{z}) + \log p(\mathbf{z}|c) \\ & + \log p(c) - \log q(\mathbf{z}|\mathbf{y}) - \log q(c|\mathbf{z})] \\ =& E_{q(\mathbf{z},c|\mathbf{y})}[\log p(\mathbf{y}|\mathbf{z}) \\ & - \log \frac{q(\mathbf{z}|\mathbf{y})}{p(\mathbf{z}|c)} - \log \frac{q(c|\mathbf{z})}{p(c)}] \\ =& E_{q(\mathbf{z}|\mathbf{y})}[\log p(\mathbf{y}|\mathbf{z}) \\ & - \sum_{c} q(c|\mathbf{z}) \log \frac{q(\mathbf{z}|\mathbf{y})}{p(\mathbf{z}|c)} \\ & - D_{KL}(q(c|\mathbf{z})||p(c))] \end{aligned} \tag{9}$$

where the first term in Equation 9 is the reconstruction term, which encourages the model to reconstruct the input. The second term aligns the latent vector $\mathbf{z}$ of input $\mathbf{y}$ to the latent topic representation corresponding to topic c. $q(c|\mathbf{z})$ can be regarded as the clustering result for the input comment sentence Y. The last term is used to narrow the distance between posterior topic distribution $q(c|\mathbf{z})$ and prior topic distribution $p(c)$.

In practical implementations, the prior topic categorical distribution $p(c)$ is set to uniform distribution $p(c) = \frac{1}{K}$ to prevent the posterior topic distribution $q(c|\mathbf{z})$ from collapsing, that is, all comments are clustered into one topic. $p(\mathbf{z}|c)$ is a parameterised diagonal Gaussian as follows:

$$p(\mathbf{z}|c) = \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{c}}, \text{diag}\left(\mathbf{1}\right)\right), \tag{10}$$

where $\boldsymbol{\mu}_{\boldsymbol{c}}$ is mean of Gaussian for topic c, which is also used as the latent topic representation of topic c. Inspired by VAE (Kingma and Welling 2014; Bowman et al. 2016), we employ a parameterised diagonal Gaussian as $q(\mathbf{z}|\mathbf{y})$:

$$\begin{aligned} \boldsymbol{\mu} = l_1(\mathbf{h}), \log \boldsymbol{\sigma} = l_2(\mathbf{h}), \\ q(\mathbf{z}|\mathbf{y}) = \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}, \text{diag}\left(\boldsymbol{\sigma}^2\right)\right), \end{aligned} \tag{11}$$

where $l_1(\cdot)$ and $l_2(\cdot)$ are linear transformations, $\mathbf{h}$ is obtained by the comment encoder, which contains two MLP layers with $\tanh$ activation functions. In addition, a classifier with two MLP layers is used to predict topic distribution $q(c|\mathbf{z})$. $p(\mathbf{y}|\mathbf{z})$ is modeled by the decoder, which is a one-layer MLP with $\mathrm{softmax}$ activation function.

After training, K reader-aware topic representation vectors $\{\boldsymbol{\mu}_i\}_1^K$ are obtained only from the whole reader comments corpus in the training set. And these reader-aware topics can be used to control the topic diversity of the generated comments.

Reader-Aware Saliency Detection

Reader-aware saliency detection component is designed to select the most important and reader-interested information from the news article. It conducts Bernoulli distribution estimating on each word of the news content, which indicates whether each content word is important or not. Then we can preserve the selected important words for comment generation.

Specifically, the saliency detection component first uses a BiLSTM encoder to encode title words and the last hidden vectors of two directions are used as the title representation $\mathbf{h}^{te}$. Then we use two-layer MLP with $\mathrm{sigmoid}$ activation function on the final layer to predict the selection probability for each content word x_i as follows jointly considering the title information:

$$p_\theta(\beta_i|x_i) = \mathrm{MLP}(\mathbf{h}_i^e, \mathbf{h}^{te}), \tag{12}$$

where $\mathbf{h}_i^e$ is the hidden vector obtained by BiLSTM-based news content encoder in Section 2.2. The probability β_i determines the probability (saliency) of the word be selected, and it is used to parameterize a Bernoulli distribution. Then a binary gate for each word can be obtained by sampling from the Bernoulli distribution:

$$g_i \sim Bernoulli\,(\beta_i) \tag{13}$$

Content words with $g_i = 1$ are selected as context information to conduct the comment generation. Thus, the weight of each content word in attention mechanism and the weighted source context in the backbone framework in Section 2.2 are changed as follows:

$$\begin{aligned} \hat{\alpha}_{ti} &= \frac{g_i \odot \exp(e_{ti})}{\sum_{i\prime=1}^{|X|} g_{i\prime} \odot \exp(e_{ti\prime})}, \\ \tilde{\mathbf{h}}_t^e &= \sum_i \hat{\alpha}_{ti} \mathbf{h}_i^e, \end{aligned} \tag{14}$$

where $\tilde{\mathbf{h}}_t^e$ is the selected saliency information of news content and it will be used for comment generation.

However, the sampling operation in Equation 13 is not differentiable. In order to train the reader-aware saliency detection component in an end-to-end manner, we apply Gumbel-Softmax distribution as a surrogate of Bernoulli distribution for each word selection gate (Xue, Li, and Zhang 2019). Specifically, the selection gate produces a two-element one-hot vector as follows:

$$\begin{aligned} \mathbf{g}_i = \text{one\_hot}\,(\arg\max_j p_{i,j}, j = 0, 1) \\ p_{i,0} = 1 - \beta_i, p_{i,1} = \beta_i \end{aligned} \tag{15}$$

we use the Gumbel-Softmax distribution to approximate to the one-hot vector $\mathbf{g}_i$:

$$\begin{aligned} \hat{\mathbf{g}}_i &= [\hat{p}_{i,j}]_{j=0,1} \\ \hat{p}_{i,j} &= \frac{\exp((\log(p_{i,j}) + \epsilon_j)/\tau)}{\sum_{j\prime=0}^{1} \exp((\log(p_{i,j\prime}) + \epsilon_{j\prime})/\tau)}, \end{aligned} \tag{16}$$

where ϵ_j is a random sample from Gumbel(0, 1). When temperature τ approaches 0, Gumbel-Softmax distribution approaches one-hot. And now we use $g_i = \hat{\mathbf{g}}_{i,0}$ instead of Equation 13. Via this approximation, we can train the component end-to-end with other modules. In order to encourage the saliency detection component to turn off more gates and select less words, a l_1 norm term over all gates is added to the loss function as follows:

$$\mathcal{L}_{sal} = \frac{\|\mathcal{G}\|_1}{|X|} = \frac{\sum_i g_i}{|X|}. \tag{17}$$

Diversified Comment Generation

Given the learned K reader-aware topic vectors, we need to select an appropriate topic for current article to guide the comment generation. Therefore, a two-layers MLP with $\mathrm{softmax}$ activation function is used to predict the selection probability of each topic as follows:

$$p(c|X) = MLP(\mathbf{h}^e). \tag{18}$$

During training, the true topic distribution $q(c|\mathbf{z})$ (Equation 8) is available and is used to compute a weighted topic representation by:

$$\tilde{\boldsymbol{\mu}} = \sum_c^K q(c|\mathbf{z}) \odot \boldsymbol{\mu_c}. \tag{19}$$

After getting the topic vector $\tilde{\boldsymbol{\mu}}$ and the selected saliency information $\tilde{\mathbf{h}}_t^e$, we update the hidden vector of the backbone decoder in Section 2.2 as follows:

$$\tilde{\mathbf{h}}_t = \tanh(W_c[\mathbf{h}_t; \tilde{\mathbf{h}}_t^e; \tilde{\boldsymbol{\mu}}]). \tag{20}$$

Then $\tilde{\mathbf{h}}_t$ is used to predict next word as Equation 5.

In the inference stage, $p(c|X)$ is used to get the topic representation. So in order to learn $p(c|X)$ during the training stage, a KL term $\mathcal{L}_{top} = D_{KL}(q(c|\mathbf{z})||p(c|X))$ is added into the final loss function.

Learning

Finally, considering all components the loss function of the whole comment generation framework is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{ELBO}} + \lambda_2 \mathcal{L}_{sal} + \lambda_3 \mathcal{L}_{ce} + \lambda_4 \mathcal{L}_{top}. \tag{21}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are hyperparameters to make a trade-off among different components. Then we jointly train all components according to Equation 21.

Experimental Settings

Datasets

Tencent Corpus is a Chinese dataset published in (Qin et al. 2018). The dataset is built from Tencent News[1] and each data item contains a news article and the corresponding comments. Each article is made up of a title and a body. All text

[1] https://news.qq.com/

		Train	Dev	Test
Tencent	# News	191,502	5,000	1,610
	Avg. # Cmts per News	5	27	27
Yahoo	# News	152,355	5,000	3,160
	Avg. # Cmts per News	7.7	20.5	20.5
NetEase	# News	75,287	5,000	2,500
	Avg. # Cmts per News	22.7	22.5	22.5

Table 1: Statistics of the three datasets.

is tokenized by a Chinese word segmenter JieBa[2]. The average lengths of news titles, news bodies, and comments are 15 words, 554 words, and 17 words respectively.

Yahoo! News Corpus is an English dataset published in (Yang et al. 2019), which is built from Yahoo! News[3]. Text is tokenized by Stanford CoreNLP (Manning et al. 2014). The average lengths of news titles, news bodies, and comments are 12, 578, and 32 respectively.

NetEase News Corpus is also a Chinese dataset crawled from NetEase News[4] and used in (Zheng et al. 2017). We process the raw data according to the processing methods used in the first two datasets (Qin et al. 2018; Yang et al. 2019). On average, news titles, news bodies, and comments contain 12, 682, and 23 words respectively.

Table 1 summarizes the statistics of the three datasets.

Baseline Models

The following models are selected as baselines:

Seq2seq (Qin et al. 2018): this model follows the framework of seq2seq model with attention. We use two kinds of input, the title(**T**) and the title together with the content (**TC**).

GANN (Zheng et al. 2017): the author proposes a gated attention neural network, which is similar to **Seq2seq-T** and adds a gate layer between encoder and decoder.

Self-attention (Chen et al. 2018): this model also follows seq2seq framework and use multi-layer self multi-head attention as the encoder and a RNN decoder with attention is applied. We follow the setting of Li et al.(2019) and use the bag of words as input. Specifically the words with top 600 term frequency are as the input.

CVAE (Zhao, Zhao, and Eskenazi 2017): this model uses conditional VAE to improve the diversity of neural dialog. We use this model as a baseline for evaluating the diversity of comments.

Evaluation Metrics

Automatic Evaluation Following Qin et al. (2018), we use ROUGE (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and METEOR (Banerjee and Lavie 2005) as metrics to evaluate the performance of different models. A popular NLG evaluation tool nlg-eval[5] is used to compute these metrics. Besides the overlapping based metrics, we use Distinct (Li et al. 2016) to evaluate the diversity of comments. Distinct-n measures the percentage of distinct n-grams in all generated results. M-Distinct-n measures the ability to generate multiple diverse comments for the same test article. For computing M-Distinct, 5 generated comments for each test article are used. For Seq2seq-T, Seq2seq-TC, GANN and Self-attention, top 5 comments from beam search with a size of 5 are used. For CVAE, we decode 5 times by sampling on latent variable to get 5 comments. For our method, we decode 5 times with one of top 5 predicted topics to get 5 comments.

Human Evaluation. Following Qin et al. (2018), we also evaluate our method by human evaluation. Given titles and bodies of news articles, raters are asked to rate the comments on three dimensions: **Relevance**, **Informativeness**, and **Fluency**. **Relevance** measures whether the comment is about the main story of the news, one side part of the news, or irrelevant to the news. **Informativeness** evaluates how much concrete information the comment contains. It measures whether the comment involves a specific aspect of some character or event. **Fluency** evaluates whether the sentence is fluent. It mainly measures whether the sentence follows the grammar. The score of each aspect ranges from 1 to 5. In our experiment, we randomly sample 100 articles from the test set for each dataset and ask three raters to judge the quality of the comments given by different models. For every article, comments from all models are pooled, randomly shuffled, and presented to the raters.

Implementation Details

For each dataset, we use a vocabulary with the top 30k frequent words in the entire data. We limit maximum lengths of news titles, news bodies and comments to 30, 600 and 50 respectively. The part exceeding the maximum length is truncated. The embedding size is set to 256. The word embedding are shared between encoder and decoder. For RNN based encoder, we use a two-layer BiLSTM with hidden size 128. We use a two-layer LSTM with hidden size 256 as decoder. For self multi-head attention encoder, we use 4 heads and two layers. For CVAE and our topic modeling component, we set the size of latent variable to 64. For our method, λ_1 and λ_3 are set to 1, λ_2 and λ_4 are set to 0.5×10^{-3} and 0.2 respectively. We choose topic number K from set [10, 100, 1000], and we set $K = 100$ for Tencent dataset and $K = 1000$ for other two datasets. The dropout layer is inserted after LSTM layers of decoder and the dropout rate is set to 0.1 for regularization. The batch size is set to 128. We train the model using Adam (Kingma and Ba 2014) with learning rate 0.0005. We also clamp gradient values into the range $[-8.0, 8.0]$ to avoid the exploding gradient problem (Pascanu, Mikolov, and Bengio 2013). In decoding, top 1 comment from beam search with a size of 5 is selected for evaluation.

Experimental Results and Discussions

Automatic and Human Evaluation

Automatic evaluation results on three datasets are shown in Table 2. On most automatic metrics, our method outperforms baseline methods. Compared with Seq2seq-TC,

[2]https://github.com/fxsjy/jieba

[3]https://news.yahoo.com/

[4]https://news.163.com/

[5]https://github.com/Maluuba/nlg-eval

Dataset	Models	ROUGE_L	CIDEr	METEOR	Distinct-3	Distinct-4	M-Distinct-3	M-Distinct-4
Tencent	Seq2seq-T	0.261	0.015	0.076	0.088	0.079	0.046	0.051
	Seq2seq-TC	0.280	0.021	0.088	0.121	0.122	0.045	0.054
	GANN	0.267	0.017	0.081	0.087	0.081	0.040	0.046
	Self-attention	0.280	0.019	0.092	0.117	0.121	0.043	0.051
	CVAE	0.281	0.021	0.094	0.135	0.137	0.041	0.044
	Ours	**0.289**	**0.024**	**0.107**	**0.176**	**0.196**	**0.092**	**0.112**
Yahoo	Seq2seq-T	0.299	0.031	0.105	0.137	0.168	0.044	0.063
	Seq2seq-TC	0.308	**0.037**	0.106	**0.179**	0.217	0.056	0.078
	GANN	0.301	0.029	0.104	0.116	0.148	0.034	0.049
	Self-attention	0.296	0.025	0.096	0.150	0.181	0.049	0.068
	CVAE	0.300	0.031	0.107	0.159	0.192	0.049	0.069
	Ours	**0.309**	0.033	**0.111**	0.169	**0.220**	**0.068**	**0.097**
NetEase	Seq2seq-T	0.263	0.025	0.105	0.149	0.169	0.046	0.056
	Seq2seq-TC	0.268	**0.035**	0.108	0.178	0.203	0.053	0.064
	GANN	0.258	0.022	0.105	0.129	0.146	0.042	0.051
	Self-attention	0.265	0.034	0.110	0.174	0.204	0.053	0.067
	CVAE	0.261	0.026	0.106	0.120	0.135	0.041	0.049
	Ours	**0.269**	0.034	**0.111**	**0.189**	**0.225**	**0.081**	**0.103**

Table 2: Automatic evaluation results on three datasets

Dataset	Models	Relevance	Informativeness	Fluency	Total
Tencent	Seq2seq-TC	1.22	1.11	**3.52**	1.95
	Self-attention	1.48	1.41	**3.52**	2.14
	CVAE	1.58	1.44	3.47	2.16
	Ours	**2.02**	**1.84**	3.49	**2.45**
Yahoo	Seq2seq-TC	1.70	1.70	3.77	2.39
	Self-attention	1.71	1.72	**3.84**	2.42
	CVAE	1.63	1.65	3.79	2.36
	Ours	**2.00**	**2.01**	3.71	**2.57**
NetEase	Seq2seq-TC	1.97	1.99	4.03	2.66
	Self-attention	1.90	1.96	4.02	2.63
	CVAE	1.50	1.53	**4.25**	2.42
	Ours	**2.10**	**2.15**	4.05	**2.76**

Table 3: Human evaluation results on three datasets

Metrics	Distinct-3	Distinct-4	M-Distinct-3	M-Distinct-4
No Topic Modeling	0.142	0.151	0.050	0.060
No Saliency Detection	0.173	0.188	0.087	0.104
Full Model	0.176	0.196	0.092	0.112

Table 4: Model ablation results on Tencent dataset

Seq2seq-T and GANN perform worse in all metrics. This indicates that news bodies are important for generating better comments. The results of Self-attention and CVAE are not stable. Compared with Seq2seq-TC, Self-attention performs worse in Yahoo dataset and close in other datasets. CVAE performs better in Tencent dataset and worse in other dastasets compared with Seq2seq-TC. Compared with other methods, our method improves Distinct-4 and M-Distinct scores significantly. This demonstrates that our method can generate diversified comments according to different topics and salient information. While different comments of one article of Seq2seq-T, Seq2seq-TC, GANN and Self-attention come from the same beam, the M-Distinct scores of these methods are lower than ours. Although CVAE can generate different comments for one article by sampling on a latent variable, it obtains a worse M-Distinct score than ours. This demonstrates that the semantic change of generated comments is small when sampling on a latent variable. Our method generates comments by selecting different topic representation vectors and salient information of news, thus has a higher M-Distinct score.

Table 3 reports human evaluation results in three datasets. Because Seq2seq-T and GANN are not using news bodies and perform worse in automatic metrics, we compare our method with other methods. Our method achieves the best Total scores in three datasets. Specifically, our method mainly improves scores on Relevance and Informativeness. This shows that our method can generate more relevant and informative comments by utilizing reader-aware topic modeling and saliency information detection. However, our method performs worse in term of Fluency. We find that baselines tend to generate more generic responses, such as "Me too.", resulting in higher Fluency scores.

Ablation Study

We conduct ablation study to evaluate the affection of each component and show the results in Table 4. We compare our full model with two variants: (1) No Topic Modeling: the reader-aware topic modeling component is removed; (2) No Saliency Detection: the reader-aware saliency detection is removed. We can see that our full model obtains the best performance and two components contribute to the performance. No topic modeling drops a lot in Distinct and M-Distinct. This shows that the reader-aware topic modeling component is important for generating diversified comments. With saliency detection, the performance gets better and this indicates that it is useful to detect important information of news for generating diversified comments.

Analysis of Learned Latent Topics

In order to evaluate the reader-aware topic modeling component, we visualize the learned latent semantic vectors of comments. To this end, we use t-SNE (Maaten and Hinton 2008) to reduce the dimensionality of the latent vector $\mathbf{z}$ to 2. The one with the highest probability in topic distribution $q(c|\mathbf{z})$ is used as the topic of a comment. We first randomly select 10 topics from 100 topics in Tencent dataset and then plot 5000 sampled comments belonging to these 10 topics in Figure 3. Points with different colors belong to different topics. In addition, we plot topic vectors of these 10 topics. We

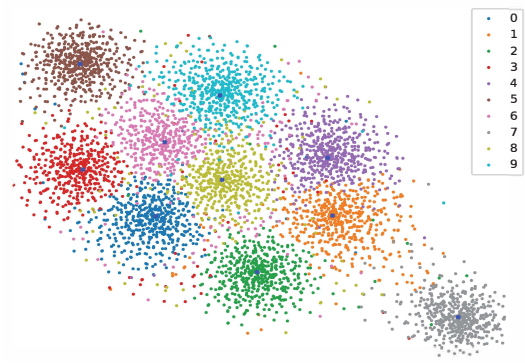

Figure 3: The latent semantic vectors of sampled comments and corresponding topic vectors.

Topic 1: 不错, 有点, 真的, 太, 挺, 比较, 应该, 确实, 好像, 其实 (nice, a little, really, too, quite, comparatively, should, indeed, like, actually)
Topic 22: 恶心, 丑, 可爱, 真, 不要脸, 太, 讨厌, 难看, 脸, 臭 (disgusting, ugly, cute, real, shameful, too, nasty, ugly, face, stink)
Topic 37: 好看, 漂亮, 挺, 不错, 身材, 演技, 颜值, 性感, 长得, 很漂亮 (good-looking, beautiful, quite, nice, body, acting, face, sexy, look like, very beautiful)
Topic 62: 好吃, 东西, 喝, 吃, 味道, 鱼, 肉, 水, 菜, 不吃 (delicious, food, drink, eat, taste, fish, meat, water, dish, don't eat)
Topic 99: 穿, 腿, 长, 衣服, 眼睛, 好看, 脸, 胖, 瘦, 高 (wear, leg, long, clothes, eye, good-looking, face, fat, thin, tall)

Table 5: The top 10 frequent words of some topics

can see that these latent vectors of comments are grouped into several clusters and distributed around corresponding topic vectors. This shows that reader-aware topic modeling component can effectively cluster comments and topic vectors can be used to represent topics well. What's more, we collect comments on each topic to observe what each topic is talking about. The top 10 frequent words (removing stop words) of some topics are shown in Table 5. We can see that Topic 1 is talking about intensity of emotional expression, such as "a little", "really", and "too". Appearance and Food are discussed in Topic 37 and Topic 62 respectively.

Case Study

In order to further understand our model, we compare comments generated by different models in Table 6. The news article is about the dressing of a female star. Seq2seq-TC produces a general comment while Self-attention produces an informative comment. However, they can not produce multiple comments for one article. CVAE can achieve this by sampling on a latent variable, but it produces same comments for different samples. Compared to CVAE, our model generates more relevant and diverse comments according to different topics. For example, "It is good-looking in any clothes" comments on the main story of news and mentions detail of news, "wearing clothes". What's more, comparing the generated comment conditioning on a specific topic with corresponding topic words in Table 5, we find that the generated comments are consistent with the semantics of the

Title: 蒋欣终于穿对衣服！尤其这开叉裙，显瘦20斤！微胖界穿搭典范！ (Jiang Xin is finally wearing the right clothes! Especially this open skirt, which is 20 pounds slimmer! Micro-fat dress code!)	
Body (truncated): 中国全民追星热的当下,明星的一举一动,以及穿着服饰,都极大受到粉丝的追捧。事实上,每位女明星都有自己的长处、优点,善意看待就好噢。蒋欣呀蒋欣,你这样穿确实很显瘦的说,着实的好看又吸睛,难怪人人都说,瘦瘦瘦也可以凹出美腻感的节奏有么有,学会这样穿说不定你也可以一路美美美的节奏。 (At the moment when China's people are star-fighting, the celebrity's every move, as well as wearing clothing, are greatly sought after by fans. In fact, each female star has its own strengths and advantages, just look at it in good faith. Jiang Xin, Jiang Xin, you are indeed very thin when you wear it.It is really beautiful and eye-catching. No wonder everyone says that there is a rhythm of thinness.You can learn to wear it like this. You can have a beautiful rhythm.)	
Seq2seq-TC: 我也是 (Me too.) **Self-attention:** 喜欢蒋欣 (Like Jiang Xin.)	
CVAE	**1:** 我喜欢蒋欣 (I like Jiang Xin.) **2:** 我喜欢蒋欣 (I like Jiang Xin.)
Ours	**Topic 99:** 穿什么衣服都好看 (It is good-looking in any clothes.) **Topic 22:** 好可爱 (So cute.) **Topic 37:** 挺好看的(It is pretty beautiful.) **Topic 1:** 不错不错 (It is nice.) **Topic 62:** 胖了 (Gain weight.)

Table 6: A Case from Tencent News dataset

corresponding topics.

Related Work

Automatic comment generation is proposed by Zheng et al.(2017) and Qin et al.(2018). The former proposed to generate one comment only based on the news title while the latter extended the work to generate a comment jointly considering the news title and body content. These two methods adopted seq2seq (Sutskever, Vinyals, and Le 2014) framework and conducted saliency detection directly via attention modeling. Recently, Li et al.(2019) extracted keywords as saliency information and Yang et al.(2019) proposed a reading network to select important spans from the news article. Huang et al.(2020) employed the LDA (Blei, Ng, and Jordan 2003) topic model to conduct information mining from the content. All these works concern the saliency information extraction. However, they neglect various reader-aware factors implied in the comments. Our method simultaneously considers these two aspects and utilizes two novel components to generate diversified comments.

Conclusion

We propose a reader-aware topic modeling and saliency detection framework to enhance the quality of generated comments. We design a variational generative clustering algorithm for topic mining from reader comments. We introduce Bernoulli distribution estimating on news content to select saliency information. Results show that our framework outperforms existing baseline methods in terms of automatic metrics and human evaluation.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 61773229 and 6201101015), Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202032), Shenzhen Giiso Information Technology Co. Ltd., the Basic Research Fund of Shenzhen City (Grand No. JCYJ20190813165003837), and Overseas Cooperation Research Fund of Graduate School at Shenzhen, Tsinghua University (Grant No. HW2018002).

Ethics Statement

We are fully aware of the new ethics policy posted in the AAAI 2021 CFP page and we seriously honor the AAAI Publications Ethics and Malpractice Statement, as well as the AAAI Code of Professional Conduct. And along the whole process of our research project, we carefully think about them. Here we elaborate on the ethical impact of this task and our method.

Automatic comment generation aims to generate comments for news articles. This task has many potential applications. First, researchers have been working to develop a more intelligent chatbot (such as XiaoIce (Shum, He, and Li 2018; Zhou et al. 2020)), which can not only chat with people, but also write poems, sing songs and so on. The application of this task is to give the chatbot the ability to comment on articles (Qin et al. 2018; Yang et al. 2019) to enable in-depth, content-rich conversations with users based on articles users are interested in. Second, it can be used for online discussion forums to increase user engagement and foster online communities by generating some enlightening comments to attract users to give their own comments. Third, we can build a comment writing assistant which generates some candidate comments for users (Zheng et al. 2017). Users could select one and refine it, which makes the procedure more user-friendly. Therefore, this task is novel and meaningful.

We are aware that numerous uses of these techniques can pose ethical issues. For example, there is a risk that people and organizations could use these techniques at scale to feign comments coming from people for purposes of political manipulation or persuasion (Yang et al. 2019). Therefore, in order to avoid potential risks, best practices will be necessary for guiding applications and we need to supervise all aspects when deploying such a system. First, we suggest that market regulators must monitor organizations or individuals that provide such services to a large number of users. Second, we suggest limiting the domain of such systems and excluding the political domain. And some post-processing techniques are need to filter some sensitive comments. Third, we suggest limiting the number of system calls in a short period of time to prevent massive abuse. We believe that reasonable guidance and supervision can largely avoid these risks.

On the other hand, we have to mention that some typical tasks also have potential risks. For example, the technology of dialogue generation (Zhang et al. 2020) can be used to disguise as a normal person to deceive people who chat with it, and achieve a certain purpose. The technology of face generation (Karras et al. 2018) can be used to disguise as the face of target people, so as to deceive the face recognition system. However, there are still many researchers working on these tasks for the positive uses of these tasks. Therefore, everything has two sides and we should treat it dialectically.

In addition, we believe that the study of this technology is important for us to better understand the defects of this technology, which helps us to detect spam comments and combat this behavior. For example, Zellers et al.(2019) found that the best defense against fake news turns out to be a strong fake news generator.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan): 993–1022.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 10–21. Association for Computational Linguistics.

Chen, M. X.; Firat, O.; Bapna, A.; Johnson, M.; Macherey, W.; Foster, G.; Jones, L.; Schuster, M.; Shazeer, N.; Parmar, N.; et al. 2018. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 76–86. Association for Computational Linguistics.

Gatt, A.; and Krahmer, E. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61: 65–170.

Huang, J.; Pan, L.; Xu, K.; Peng, W.; and Li, F. 2020. Generating Pertinent and Diversified Comments with Topic-aware Pointer-Generator Networks. *arXiv preprint arXiv:2005.04396* .

Jiang, Z.; Zheng, Y.; Tan, H.; Tang, B.; and Zhou, H. 2017. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1965–1972.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

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

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *stat* 1050: 1.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003. Association for Computational Linguistics.

Li, W.; Xu, J.; He, Y.; Yan, S.; Wu, Y.; and Sun, X. 2019. Coherent Comments Generation for Chinese Articles with a Graph-to-Sequence Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4843–4852. Florence, Italy: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 1310–1318.

Qin, L.; Liu, L.; Bi, W.; Wang, Y.; Liu, X.; Hu, Z.; Zhao, H.; and Shi, S. 2018. Automatic Article Commenting: the Task and Dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 151–156. Melbourne, Australia: Association for Computational Linguistics.

Reiter, E.; and Dale, R. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3(1): 57–87.

Shum, H.-Y.; He, X.-d.; and Li, D. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19(1): 10–26.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Xue, L.; Li, X.; and Zhang, N. L. 2019. Not All Attention Is Needed: Gated Attention Network for Sequence Data. *arXiv preprint arXiv:1912.00349* .

Yang, Z.; Xu, C.; Wu, W.; and Li, Z. 2019. Read, Attend and Comment: A Deep Architecture for Automatic News Comment Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5077–5089. Hong Kong, China: Association for Computational Linguistics.

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, 9054–9065.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278. Online: Association for Computational Linguistics.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 654–664.

Zheng, H.-T.; Wang, W.; Chen, W.; and Sangaiah, A. K. 2017. Automatic generation of news comments based on gated attention neural networks. *IEEE Access* 6: 702–710.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46(1): 53–93.