Aspect-Level Sentiment-Controllable Review Generation with Mutual Learning Framework

Huimin Chen,[12] Yankai Lin,[3] Fanchao Qi,[2] Jinyi Hu,[2] Peng Li,[3] Jie Zhou,[3] Maosong Sun[2*]

[1]School of Journalism and Communication, Tsinghua University, Beijing, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China
[3]Pattern Recognition Center, WeChat AI, Tencent Inc., China
huimchen@mail.tsinghua.edu.cn

Abstract

Review generation, aiming to automatically generate review text according to the given information, is proposed to assist in the unappealing review writing. However, most of existing methods only consider the overall sentiments of reviews and cannot achieve aspect-level sentiment control. Even though some previous studies attempt to generate aspect-level sentiment-controllable reviews, they usually require large-scale human annotations which are unavailable in the real world. To address this issue, we propose a mutual learning framework to take advantage of unlabeled data to assist the aspect-level sentiment-controllable review generation. The framework consists of a generator and a classifier which utilize confidence mechanism and reconstruction reward to enhance each other. Experimental results show our model can achieve aspect-sentiment control accuracy up to 88% without losing generation quality.

Introduction

With the rapid development of the Internet, more and more websites providing online review service appear (e.g., Amazon, Yelp and TripAdvisor). Online reviews can not only help customers with their purchase decisions, but also contribute to the online platforms for better product recommendations (McAuley and Leskovec 2013). Unfortunately, most of the customers only give a rating (Chen and Xie 2008) rather than write a paragraph of review text, since it is not so pleasant to write down opinions for most people. To assist the procedure of review writing and make it more efficient and user-friendly, Lipton, Vikram, and McAuley (2015) first propose the task of review generation, which aims to automatically generate a review when given the customer's overall sentiment (rating) towards a product as input.

In recent years, various review generation models have been proposed (Lipton, Vikram, and McAuley 2015; Tang et al. 2016; Dong et al. 2017; Li and Tuzhilin 2019). These models usually regard review generation as a text generation task conditioned on the information about user, product, and overall sentiment. Nevertheless, review writing in real-world applications is much more complicated. Users usually have different sentiments towards various aspects of a product, and therefore their reviews will cover multiple aspects of the product. For example, as shown in Figure 1, a user who gives a neutral overall rating can express positive sentiments to the *environment* and *service*, while writing negative comment towards the *food*. Moreover, reviews covering aspect-level information are more helpful when serving as references for customers' purchase decisions. Consequently, it is of considerable importance to generate aspect-level sentiment-controllable reviews.

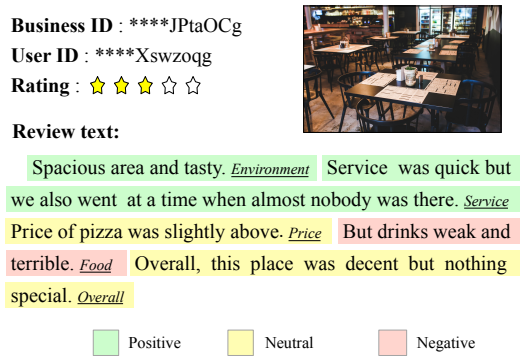

Figure 1: A review with diverse sentiments towards different aspects of a restaurant.

Zang and Wan (2017) first introduce the aspect-sentiment information to perform aspect-level sentiment-controllable review generation. They adopt a supervised method requiring a large amount of training instances with sentence-level aspect-sentiment annotations. However, very few reviews have sentence-level aspect-sentiment labels, and it is also labor-intensive and time-consuming to annotate these information for all reviews.

To tackle this issue, we propose a semi-supervised aspect-level sentiment-controllable review generation method (ASRG), which can take advantage of large-scale unlabeled data to achieve aspect-level sentiment control in review genenration with a few labeled data. We propose a mutual

*Corresponding author: M.Sun(sms@mail.tsinghua.edu.cn)

learning framework which trains a review generator and a sentiment classifier leveraging unlabeled data in a more effective way. In the mutual learning framework, the unlabeled review data plays a similar role to the labeled data to help model training. Specifically, our mutual learning framework can be described as: (1) the review generator learns from the unlabeled data using pseudo aspect-sentiment labels provided by the sentiment classifier and uses a confidence mechanism to reduce the influence of those noisy labels; (2) the sentiment classifier learns from the unlabeled data using the review reconstruction of review generator as a reinforcement learning task. Through multi-steps of mutual learning, we obtain an enhanced review generator which could generate high-quality aspect-level sentiment-controllable reviews.

We conduct evaluations on several real-world datasets, finding that our model can generate reviews with accurate aspect-level sentiment control (above 88% accuracy), while maintaining comparable review quality with state-of-the-art models. We also conduct further analyses to demonstrate the effectiveness of the confidence mechanism and the reconstruction reward.

Related Work

Automatic review generation has been recently proposed to assist the review writing and recommendation systems. Researchers have made great efforts to generate online reviews with good quality. Lipton, Vikram, and McAuley (2015) first adopt Recurrent Neural Network to generate reviews. Afterwards, attention mechanism (Bahdanau, Cho, and Bengio 2015) is introduced into review generation to condition on inputs selectively (Tang et al. 2016; Dong et al. 2017). Owing to the correlation of recommendation task and review generation task, a series of researches jointly model these two tasks to generate personalized reviews (Wang and Zhang 2017; Ni et al. 2017; Ni, Li, and McAuley 2019). As the aspect information contributes to the review generation, Ni and McAuley (2018) and Li et al. (2019) focus on taking aspect information into consideration to improve the quality. Despite the notable progress, they all ignore the aspect-level sentiment control which is practical in review generation. Only Zang and Wan (2017) attempt to generate reviews from aspect-sentiment scores, which require the reviews with sentence-level aspect-sentiment score annotations. This makes it impractical in real-world applications due to the lack of labeled data. Different from them, we propose a mutual learning framework which utilizes both small-scale labeled data and large-scale unlabeled data to enhance the aspect-level sentiment-controllable review generation.

Our work is also related to low-resource text generation, which is one of the recent spotlights in NLP. Researchers have explored low-resource text generation in various applications such as machine translation (Cheng et al. 2016; Gu et al. 2018), headline generation (Tilk and Alumäe 2017), dialog generation (Tran and Nguyen 2018), data-to-text generation (Ma et al. 2019), and poem generation (Chen et al. 2019). Qader, Portet, and Labbé (2019) first introduce joint learning of natural language generation and natural language understanding models to tackle the low-resource text generation. Different from them, we make the first effort to explore the mutual learning framework in fine-grained aspect-level sentiment-controllable review generation. Furthermore, we introduce the confidence mechanism as well as constrained reconstruction reward to alleviate the noises brought from unlabeled data, while they neglect the noises.

Another closely related line of research is sentiment-controllable text generation (Hu et al. 2017; Cagan, Frank, and Tsarfaty 2017; Wang and Wan 2018; Li et al. 2020). Similar to them, our work also aims to generate text with controllable sentiments. However, we focus on utilizing the unlabeled data to improve the performance of the generator with mutual learning framework, while they focus more on how to disentangle different attributes, such as content and sentiment, to achieve the control of sentiment.

Method

In this section, we introduce our semi-supervised aspect-level sentiment-controllable review generation method (ASRG), which utilizes a mutual learning framework to learn a review generator and a sentiment classifier from both labeled and unlabeled data.

We first give the formalization of the aspect-level sentiment-controllable review generation task. Formally, given a user u, a product p, an overall sentiment s, and a list of aspect labels $a = \{a_1, a_2, ..., a_n\}$ together with their sentiment labels $y = \{y_1, y_2, ..., y_n\}$, the task aims to generate a review x comprising n sentences $x = \{x_1, x_2, \cdots, x_n\}$, each of which x_i describes the aspect a_i with the specified sentiment y_i.

In this task, we have labeled reviews L and unlabeled reviews V. Each labeled review $l \subseteq L$ comprises review text and attributes including the user, product, and overall sentiment, as well as aspect and sentiment labels for each sentence in the review text: $l =< u, p, s, a, y, x >$, while each unlabeled review $v \subseteq V$ only contains review text and attributes including the user, product, and overall sentiment: $v =< u, p, s, x >$.

In the following, we first introduce the overall mutual learning framework. Afterwards, we describe the review generator and sentiment classifier.

Mutual Learning Framework

Our ASRG model consists of a review generator G and a sentiment classifier C. The generator G is used to generate a review according to specified attributes including user, product, overall sentiment, a list of aspects together with corresponding sentiments. The classifier C is supposed to predict the aspect and sentiment of each sentence in a review. Intuitively, the generator and classifier can improve each other with the help of extra unlabeled data.

Specifically, as shown in Figure 2, we design a four-step mutual learning procedure:

Step 0: We use the small amounts of labeled data to train a weak generator G_0 and a weak classifier C_0 separately.

Step 1: Supposing the classifier C_0 can predict relatively accurate aspect and sentiment labels for given reviews, we use C_0 to classify the huge amounts of unlabeled data and obtain predicted aspect and sentiment labels first. Then we train the

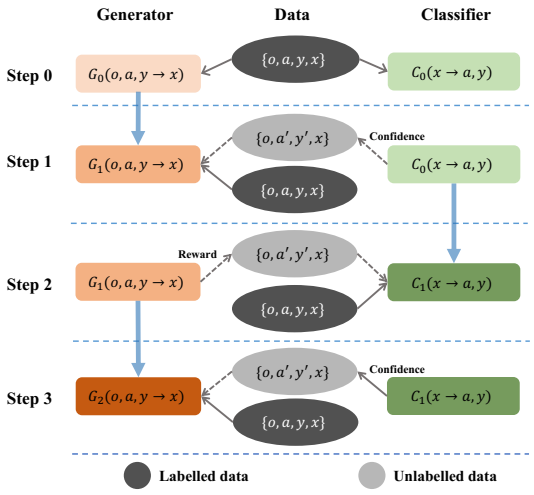

Figure 2: The overall training process of our mutual learning framework. o represents the outline information including user, product, and overall sentiment of a review. a and y denote the true aspect and sentiment labels of each sentence in a review, while a' and y' are pseudo labels predicted by the classifier C.

generator G_0 again with both labeled and unlabeled data to get the enhanced generator G_1, where the prediction results of the unlabeled data from the classifier C_0 are regarded as pseudo labels with an extra confidence mechanism.

Step 2: Supposing the generator G_1 can generate correct reviews for given aspect and corresponding sentiment labels, we use the generator G_1 to calculate generative probability for each unlabeled review given the aspect and sentiment labels predicted by the classifier C_0, and then treat the generative probability as a reconstruction reward to train the classifier C_0 for the unlabeled review. In this fashion, we obtain an improved classifier C_1.

Step 3: With the improved classifier C_1, we can repeat Step 1 to enhance the generator G_1 and obtain G_2.

Finally, we obtain the desired generator G_2 which has been enhanced by the improved classifier using the unlabeled data.

Review Generator

We introduce the architecture of our review generator first, and then describe the loss functions of labeled data and unlabeled data.

Architecture We adopt the well-established encoder-decoder architecture (Sutskever, Vinyals, and Le 2014) in our generator, which generates the whole review sentence by sentence. When generating a sentence of a review, the inputs to the generator can be divided into two parts: one part is the user, product, and overall sentiment of the review, as well as generated context, which we call "overall information"; the other part is the aspect and corresponding sentiment of the sentence, namely "aspect-sentiment information".

Specifically, when generating the i-th sentence x_i, we obtain the representation of the overall information o_i through an MLP layer:

$$\boldsymbol{o}_i = \text{MLP}([\boldsymbol{u}, \boldsymbol{p}, \boldsymbol{s}, \boldsymbol{c}_i]), \tag{1}$$

where $\boldsymbol{u}$, $\boldsymbol{p}$ and $\boldsymbol{s}$ are the embeddings of user, product, and overall sentiment, respectively, and $\boldsymbol{c}_i$ is the representation of previous $i-1$ sentences. $\boldsymbol{c}_i$ is encoded by a MLP layer and a convolution layer:

$$\begin{aligned} \boldsymbol{c}_i &= \text{MLP}(\boldsymbol{c}_{i-1}, \boldsymbol{x}_{i-1}), \\ \boldsymbol{x}_{i-1} &= \text{Conv}([\boldsymbol{x}_{i-1,1}, \cdots, \boldsymbol{x}_{i-1,m}]), \end{aligned} \tag{2}$$

where $\boldsymbol{x}_{i-1}$ represents the representation of $i-1$-th sentence obtained by the convolution layer, and m is the number of words in x_{i-1}. For the representation of aspect-sentiment information, inspired from text generation with variational autoencoder (Bowman et al. 2016; Zhao, Zhao, and Eskenazi 2017), we obtain the joint aspect-sentiment representation $\boldsymbol{z}_i$ by sampling from a multivariate Gaussian distribution: $\mathcal{N}(\boldsymbol{\mu}_{a_i,y_i}, \boldsymbol{\Sigma}_{a_i,y_i})$, to enhance the diversity of generated text for each aspect-sentiment input pair. Mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are randomly initialized for each aspect-sentiment pair.

Afterwards, we use GRU (Cho et al. 2014) as the decoder to generate the sentence x_i, with the representations of overall information $\boldsymbol{o}_i$ and aspect-sentiment representation $\boldsymbol{z}_i$ as inputs:

$$\begin{aligned} \boldsymbol{h}^G_{i,0} &= \text{MLP}([\boldsymbol{o}_i, \boldsymbol{z}_i]), \\ \boldsymbol{h}^G_{i,t} &= \text{GRU}(\boldsymbol{h}^G_{i,t-1}, \boldsymbol{x}_{i,t-1}), \\ \boldsymbol{p}^G_{i,t} &= \text{softmax}(W^G \boldsymbol{h}^G_{i,t}), \\ x_{i,t} &= \underset{j}{\text{argmax}}[\boldsymbol{p}^G_{i,t}]_j, \end{aligned} \tag{3}$$

where $\boldsymbol{h}^G_{i,t}$ and $\boldsymbol{p}^G_{i,t}$ are the hidden state and the generative probability distribution for the t-th word, respectively. $\boldsymbol{x}_{i,t}$ is the embedding of the t-th word $x_{i,t}$, and $[\,]_j$ denotes the j-th dimension of a vector.

Besides, to improve the control accuracy of aspect and sentiment of the generator, a prediction module is set which predicts the aspect and sentiment labels from hidden states of the decoder through an mean pooling layer and a softmax layer:

$$\begin{aligned} \bar{\boldsymbol{h}}^G_i &= \text{mean}(\boldsymbol{h}^G_{i,1}, \cdots, \boldsymbol{h}^G_{i,m}), \\ \boldsymbol{p}^{Ga}_i &= \text{softmax}(W^{Ga} \bar{h}^G_i), \\ \boldsymbol{p}^{Gy}_i &= \text{softmax}(W^{Gy} \bar{h}^G_i), \end{aligned} \tag{4}$$

where $\boldsymbol{p}^{Ga}_i$ and $\boldsymbol{p}^{Gy}_i$ denote the prediction probabilities of the aspect and the sentiment based on the sentence representation $\bar{\boldsymbol{h}}^G_i$, respectively.

Loss Function For labeled data, we train the generator with the true aspect label a_i and sentiment label y_i as inputs. The training loss is defined as:

$$\begin{aligned} L^{Gx}_l &= -\sum_i \sum_t \log[\boldsymbol{p}_{i,t}]_{I^x(x_{i,t})}, \\ L^{Ga}_l &= -\sum_i \log[\boldsymbol{p}^{Ga}_i]_{I^a(a_i)}, \\ L^{Gy}_l &= -\sum_i \log[\boldsymbol{p}^{Gy}_i]_{I^y(y_i)}, \\ L^G_l &= L^{Gx}_l + L^{Ga}_l + L^{Gy}_l. \end{aligned} \tag{5}$$

where L_l^{Gx} denotes the loss of generation quality, L_l^{Ga} and L_l^{Gy} are losses of the aspect prediction and corresponding sentiment prediction, respectively. $I^x()$ is the index function of a vector.

For unlabeled data, we take the pseudo aspect label a_i' and sentiment label y_i' predicted by the classifier C as inputs. Considering the noises brought from the pseudo labels, we introduce a confidence mechanism to the generator, with the predicted probabilities of pseudo labels in the classifier as confidence values.

Specifically, the impact of noisy labels mainly reflects on two parts: one is the quality of generation, another is the control of aspect and corresponding sentiment. As for generation quality, we define the loss function as L_v^{Gx} to reduce the impact of noise:

$$\begin{aligned} L_v^{Gx} = &- \sum_i \sum_t \log[\boldsymbol{p}_{i,t}]_{I^x(x_{i,t})} \cdot \\ &([\boldsymbol{p}_i^{Ca}]_{I^a(a_i')})^{\alpha} \cdot ([\boldsymbol{p}_i^{Cy}]_{I^y(y_i')})^{\beta}, \end{aligned} \tag{6}$$

where $[\boldsymbol{p}_i^{Ca}]_{I^a(a_i')}$ and $[\boldsymbol{p}_i^{Cy}]_{I^y(y_i')}$ denote the predicted probabilities of the pseudo labels a_i' and y_i' in classifier C, viewed as confidence values to weight for unlabeled data. α and β are hyper-parameters. As for aspect-sentiment control, to alleviate the noise, the loss function of this part is defined as:

$$\begin{aligned} L_v^{Ga} &= - \sum_i \log[\boldsymbol{p}_i^{Ga}]_{I^a(a_i')} \cdot ([\boldsymbol{p}_i^{Ca}]_{I^a(a_i')})^{\alpha}, \\ L_v^{Gy} &= - \sum_i \log[\boldsymbol{p}_i^{Gy}]_{I^y(y_i')} \cdot ([\boldsymbol{p}_i^{Cy}]_{I^y(y_i')})^{\beta}, \end{aligned} \tag{7}$$

where L_v^{Ga} and L_v^{Gy} represent the losses of aspect control and sentiment control, respectively. $[\boldsymbol{p}_i^{Ca}]_{I^a(a_i')}$ and $[\boldsymbol{p}_i^{Cy}]_{I^y(y_i')}$ are the same confidence values as in Eq. (6). Based on the losses of generation quality part and aspect-sentiment control part, the training loss of unlabeled data is:

$$L_v^G = L_v^{Gx} + L_v^{Ga} + L_v^{Gy}. \tag{8}$$

Therefore the review generator utilizes both labeled data and unlabeled data for training, with the labeled data trained with L_l^G and the unlabeled data trained using L_u^G.

Sentiment Classifier

We describe the architecture of our sentiment classifier first, and then loss functions of labeled data and unlabeled data.

Architecture We first use a GRU layer to encode the sentence in the review and get the hidden state in timestep t:

$$\boldsymbol{h}_{i,t}^C = \mathrm{GRU}(\boldsymbol{h}_{i,t-1}^C, \boldsymbol{x}_{i,t-1}). \tag{9}$$

Afterwards, we utilize an attention mechanism to obtain the aspect-associated sentence representation $\boldsymbol{h}_i^{Ca}$:

$$\begin{aligned} \boldsymbol{h}_i^{Ca} &= \sum_t \omega_{i,t}^a \boldsymbol{h}_{i,t}^C, \\ \omega_{i,t}^a &= \mathrm{softmax}(\boldsymbol{v}^\top \boldsymbol{h}_{i,t}^C) \end{aligned} \tag{10}$$

where $\omega_{i,t}^a$ is the attention weight of t-th word, and $\boldsymbol{v}$ is a learnable parameter vector. Through a similar way, we can get the sentiment-associated sentence representation $\boldsymbol{h}_i^{Cy}$. By feeding the two sentence representations into two perceptrons, we can obtain the aspect prediction probability $\boldsymbol{p}_i^{Ca}$ and the sentiment prediction probability $\boldsymbol{p}_i^{Cy}$:

$$\begin{aligned} \boldsymbol{p}_i^{Ca} &= \mathrm{softmax}(W^{Ca} \boldsymbol{h}_i^{Ca}), \\ \boldsymbol{p}_i^{Cy} &= \mathrm{softmax}(W^{Cy} [\boldsymbol{h}_i^{Ca}, \boldsymbol{h}_i^{Cy}]), \end{aligned} \tag{11}$$

where W^{Ca} and W^{Cy} are parameter matrics.

Loss Function For the labeled data, we directly adopt the cross-entropy loss to train the classifier:

$$L_l^C = \sum_i (-\log[\boldsymbol{p}_i^{Ca}]_{I^a(a_i)} - \log[\boldsymbol{p}_i^{Cy}]_{I^y(y_i)}), \tag{12}$$

where a_i and y_i are the true aspect and sentiment labels of the sentence.

For the unlabeled data, since true labels are unknown, we introduce a reconstruction reward to optimize the classifier, inspired by He et al. (2016) and Luo et al. (2019). We define the reconstruction reward for unlabeled sentence as the overall generative probability of generator G given the predicted aspect and sentiment labels from classifier C. Formally, the reward is:

$$R_{a_i', y_i'} = \prod_t [\boldsymbol{p}_{i,t}^G]_{I^x(x_{i,t})}, \tag{13}$$

where a_i' and y_i' are the sentence aspect and sentiment labels predicted by the classifier C, and $\boldsymbol{p}_{i,t}^G$ is the generator G's generative probability for the t-th word according to a_i' and y_i' (as defined in Eq. (3)). Considering the inaccuracy of the generator, we apply a reward threshold λ to constraint the reconstruction reward:

$$\widetilde{R}_{a_i', y_i'} = \begin{cases} 0 & R_{a_i', y_i'} < \lambda \\ R_{a_i', y_i'} & R_{a_i', y_i'} \geq \lambda. \end{cases} \tag{14}$$

The final classifier loss for the unlabeled data is:

$$L_v^C = - \sum_i \mathbb{E}_{a_i' \sim \boldsymbol{p}_i^{Ca}, y_i' \sim \boldsymbol{p}_i^{Cy}} \widetilde{R}_{a_i', y_i'}. \tag{15}$$

Hence the classifier is trained with L_u^C for unlabeled data and L_l^C for labeled data simultaneously.

Experiments

Datasets and Settings

We conduct experiments of the ASRG task on two real-world datasets: Yelp Restaurant dataset [1] and RateBeer dataset (McAuley, Leskovec, and Jurafsky 2012).

Labeled Datasets. To the best of our knowledge, there is no off-the-shell review dataset with both sentence-level aspect-sentiment labels and user, product, and overall sentiment information.[2] Hence we manually build two labeled review datasets, including $1,000$ reviews for Yelp Restaurant dataset and RateBeer dataset, respectively. Each sentence in the review is annotated into one of 6 aspect classes and 3 sentiment classes. To ensure the quality of labeling, we ask at least two annotators to annotate each sentence. If two

[1] https://www.yelp.com/dataset

[2] The dataset used by Zang and Wan (2017) lacks user, product, and overall sentiment information, which can not be applied to baseline models and our method, as well as real-world scenes.

Datasets	#Sents	Aspects						Sentiments		
Yelp	4,915	#Food 1,860	#Service 848	#Environment 462	#Price 211	#Overall 1,177	#Other 357	#Positive 2,106	#Neutral 1,072	#Negative 1,737
Ratebeer	5,118	#Look 921	#Smell 776	#Taste 1,219	#Feel 507	#Overall 986	#Other 709	#Positive 1,631	#Neutral 2,617	#Negative 870

Table 1: Statistics of the two labeled review datasets. Each sentence in a review is labeled into one of 6 aspect classes and 3 sentiment classes. Sents denotes the abbreviation of sentences.

Datasets	#Users	#Items	#Reviews	#Sents	#Words
Yelp	40,014	34,915	444,323	2,168,848	22,612
Ratebeer	6,801	23,745	1,437,537	7,441,045	44,014

Table 2: Details of the unlabeled datasets. Sents is the abbreviation of sentences.

annotators disagree on the sentence, it will be assigned to a more experienced annotator to decide the final label referring to previous annotations. We use 500 reviews in each dataset for supervised training, and 250 reviews for validation and test, respectively. Statistics of the two labeled datasets are reported in Table 1.

Unlabeled Datasets. Unlike the labeled datasets, unlabeled datasets consist of reviews with only user, product, and overall sentiment information and no aspect-sentiment labels of each sentence. Details of the unlabeled datasets are shown in Table 2.

Experimental Settings. We discard the reviews comprising more than 10 sentences and conduct tokenization with NLTK[3]. Reviews containing any sentence with more than 20 words are removed. Words occurring less than 10 times, users and products less than 5 times are also filtered out. The dimension of word embeddings is 512, and the embedding dimensions of user, product, and overall sentiment are all set to 256. In the review generator, the sizes of outline and aspect-sentiment representations are 512. The hidden states in the sentence decoder and sentiment classifier are also 512-dimensional. We tune the hyper-parameters on the validation set, and set α and β in the generator to 0.3 and 0.5, respectively. λ in the classifier is set to 0.05. Adam (Kingma and Ba 2014) is used for optimization, and the batch sizes of both the generator and classifier are 256. We also use dropout (drop rate $= 0.25$) to avoid over-fitting. Training of the generator is stopped when the perplexity on the validation set no longer decreases, with max epoch number of 20. Training of the classifier is stopped if the performance does not improve on the validation set.

Baselines

We compare our model with several state-of-the-art review generation models including: (1) **gC2S** (Tang et al. 2016), which applies a gating mechanism to the encoder-decoder framework, to control information flow from user inputs or preceding words. (2) **Attr2Seq** (Dong et al. 2017), which adopts an attention mechanism to generate reviews conditioned on the user, product, and overall sentiment attributes differently. (3) **C2F** (Li et al. 2019), which decomposes the review generation process into a coarse-to-fine pipeline, namely aspect sequence generation, sketch generation and sentence generation. (4) **HRGM** (Zang and Wan 2017), which presents a hierarchical LSTM decoder to generate long reviews from aspect-sentiment scores through a supervised framework.

Baselines including gC2S, Attr2Seq and C2F focus on generating reviews with user, product, and overall sentiment inputs, without aspect-level control of sentiments. Hence we merely compare with these three baselines with respect to the generation quality in Section . HRGM is an aspect-level sentiment-controllable baseline, but without user and product inputs, overall sentiment control as well. For fair comparison, we input the same overall information as other baselines to HRGM. We also compare our model with **GT** (Ground Truth) in manual quality evaluation part.

Review Quality Evaluation

We conduct automatic and human evaluation on quality of generated reviews and compare our model with baselines.

Automatic Evaluation. Following previous work (Li et al. 2019), we adopt BLEU (Papineni et al. 2002) and ROUGE (Lin 2004) as metrics of automatic evaluation, which compare the similarity between the generated text and ground truth based on n-gram matching. Besides, we utilize METEOR (Lavie and Agarwal 2007), a widely used metric in machine translation, which introduces WordNet stems and synonyms to enhance the consistency with human judgment.

Human Evaluation. Following (Zang and Wan 2017), we use three criteria: (1) Readabilty (denotes fluency and coherence of the review), (2) Accuracy (represents how well the review conveys the overall sentiment), (3) Usefulness (expresses whether the review provides useful information). Each criterion is scored from 1 to 5 points, where 1-point means "very terrible" and 5-point represents "very satisfying". We invite 10 human annotators familiar with the domain of restaurant reviews, and split them into 2 groups to perform evaluation independently. The scores of two groups are averaged to avoid personal bias. Overall is the average score of three criteria.

Table 3 shows the automatic evaluation results. ASRG-G_0 is our generator trained with labeled data without mutual learning process in step 0, while ASRG-G_1 and ASRG-G_2 are the generators trained in step 1 and step 3, respectively. We split results of each dataset into two parts: the top part without

[3] https://www.nltk.org

Dataset	Models	AS Control	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Yelp	gC2S	×	18.63	9.36	5.01	2.84	18.70	**12.79**
	Attr2Seq	×	15.70	7.21	3.57	1.88	15.85	10.54
	C2F	×	15.80	7.58	3.99	2.28	20.69	8.57
	HRGM	√	24.32	9.41	3.92	1.66	16.66	9.44
	ASRG-G_0	√	12.77	3.68	0.68	0.22	12.61	4.71
	ASRG-G_1	√	25.57	12.60	6.92	4.16	22.82	11.48
	ASRG-G_2	√	**27.05**	**14.07**	**8.28**	**5.17**	**24.81**	12.06
RateBeer	gC2S	×	33.54	21.03	14.34	10.10	33.33	15.46
	Attr2Seq	×	33.66	20.60	13.73	9.49	32.37	15.20
	C2F	×	31.75	16.39	9.62	5.93	28.28	12.40
	HRGM	√	18.64	1.79	3.85	1.78	18.08	10.35
	ASRG-G_0	√	3.92	1.43	0.46	0.14	17.43	4.99
	ASRG-G_1	√	33.56	20.83	14.02	9.81	33.79	16.07
	ASRG-G_2	√	**35.24**	**21.92**	**14.76**	**10.24**	**34.23**	**16.47**

Table 3: Automatic quality evaluation of generated reviews. AS Control denotes Aspect-Sentiment Control.

Models	Readability	Accuracy	Usefulness	Overall
gC2S	3.16	4.05	3.82	3.68
Attr2Seq	3.40	4.05	3.46	3.64
C2F	4.09	3.79	3.55	3.81
HRGM	3.38	3.16	2.90	3.15
ASRG-G_0	1.15	1.75	1.17	1.36
ASRG-G_1	4.07	4.08	3.82	3.99
ASRG-G_2	**4.25**	**4.23**	**3.88***	**4.12***
GT	4.39	4.41	3.96	4.25

Table 4: Human evaluation of quality on Yelp dataset. $*$ denotes significantly better than baselines ($p < 0.01$).

the ability of aspect-level sentiment control and the bottom part with the capability. From the results we can observe that: (1) Compared to baselines without aspect-level sentiment control, our ASRG-G_2 model achieves better performance than the state-of-the-art models in terms of generation quality. This not only denotes the effectiveness of our ASRG model but also shows the importance of aspect-sentiment information. (2) Compared to the aspect-level sentiment-controllable model HRGM, our ASRG-G_1 and ASRG-G_2 model achieve significant and consistent improvements over two datasets. Note that although we have pre-trained HRGM with unlabeled data for fair comparisons, HRGM still has a poor performance in generation quality. This verifies that our ASRG model more effectively exploits the unlabeled data to enhance the generation quality by the mutual learning framework. (3) From ASRG-G_0 to ASRG-G_1 to ASRG-G_2, the generation quality is improved continuously. This indicates the significance of our mutual learning framework which applies a classifier to promote the generator and a better classifier contributes to a more professional generator.

Table 4 shows results of human evaluation. Similar to the automatic evaluation results, our model performs consistently better than state-of-the-arts in all three criteria, which implies reviews generated with aspect-level sentiment information are more readable and helpful for humans. Meanwhile, the performances of ASRG-G_0, ASRG-G_1 and ASRG-G_2 are improved gradually, which also confirms the effect of mutual learning framework. However, we can find that there is still some gap between generated reviews and ground truth (GT). The Cohen's kappa coefficient is 0.55, which is acceptable considering the complexity of the task and subjectivity of the metrics may lead to personal bias.

Models	Aspect Accuracy	Sentiment Accuracy
HRGM	0.665	0.440
ASRG-G_0	0.200	0.250
ASRG-G_1	0.835	0.825
ASRG-G_2	**0.910***	**0.880***

Table 5: Aspect-Sentiment control evaluation of generated reviews on the Yelp dataset. $*$ denotes significantly better than baselines ($p < 0.001$).

Aspect-Sentiment Control Evaluation

We evaluate the accuracy of aspect-sentiment control by human evaluation in this part. We invite 5 annotators to evaluate the accuracy of generated sentences in each review, by asking the annotator whether this sentence describes the specific aspect and the corresponding sentiment. Each review is evaluated by two annotators, and the overall accuracy is reported.

Table 5 shows the control accuracy of both aspect and sentiment. We can observe that: (1) Our ASRG-G_1 and ASRG-G_2 model obtain better accuracy compared with baseline HRGM both on the control of aspect and sentiment. HRGM is in poor control because of few labeled data, while our models can take advantage of numerous unlabeled data to develop control. (2) Benefiting from the mutual learning framework, the control accuracy of both aspects and sentiments are increased progressively from ASRG-G_0 to ASRG-G_1 and then to ASRG-G_2. It confirms that the control performance can also be enhanced by a more robust classifier. The Cohen's kappa coefficient is 0.62 in this control evaluation.

Effect of Confidence Mechanism

Figure 3 shows the effect of confidence mechanism in the review generator, in both review quality and aspect-sentiment control. From the figure, we can observe that generators

Aspect-Sentiment	HRGM	ASRG-G1	ASRG-G2
Environment-Neg	**The staff was Johnny on the spot and friendly.**	**It is a little pricey for what you get.**	It's hard to find parking.
Service-Pos	The service was great and the food was amazing.	**The place is clean and the staff is friendly.**	The staff is very friendly and the food is pretty good.
Price-Neg	It 's very pricey for what it is and the music is quite loud.	The prices are a bit high for what you get.	The prices are a bit high for what you get.
Food-Pos	The food is great.	I tried the hotdog and it was pretty good.	**But** the quality of the food is good.
Environment-Pos	**It 's better to walk here.**	**The place is nice and clean.**	The place is clean and decorated.
Food-Pos	**The MGM parking employees don't know how to direct traffic.**	If you are looking for a good sandwich, this is the place to go.	They **also** have a good selection of baked goods.
Overall-Pos	Will be back !	If you are in the area, stop by.	If you are in the area, this is the place to go.

Table 6: Cases of generated reviews. The input of overall sentiment is positive. Sentences in red are in poor aspect-sentiment control. Sentences in bold black are repetitive. Words in blue connect sentences in both content and sentiment. Pos and Neg are the abbreviations of Positive and Negative.

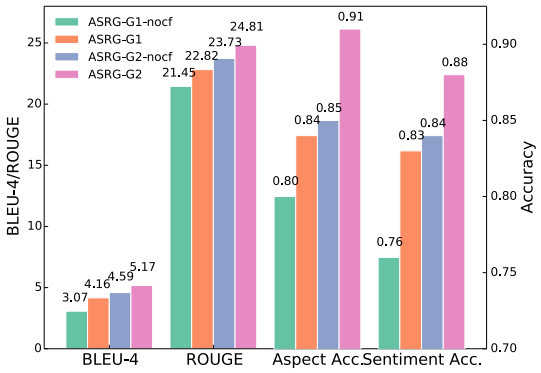

Figure 3: Effect of confidence mechanism in review quality and aspect-sentiment control . ASRG-G_1-nocf and ASRG-G_2-nocf represent generators without confidence mechanism. Acc. denotes Accuracy.

trained with confidence mechanism perform better in both review quality and aspect-sentiment control, in both training steps as well. It demonstrates that the confidence mechanism can alleviate the noises brought from the classifier and improve the generator in quality and aspect-sentiment control.

Effect of Reconstruction Reward

Table 7 shows the effect of reconstruction reward in classification accuracy. ARSG-C_0 is the classifier trained with few labeled data in step 0, with ARSG-C_1 representing the classifier strengthened with unlabeled data by reconstruction reward in step 2. From the table, we can find that ARSG-C_1 performs better than ARSG-C_0 in both datasets. It indicates the reconstruction reward obtained from the generator can effectively promote the classifier.

Case Study

Table 6 provides several cases of HRGM and ASRG models. From the table we can find that: (1) ARSG-G_1 and ARSG-G_1 outperform HRGM which generates sentences with poor aspect-sentiment control because of underutilization of unlabeled data. (2) The generator in ASRG model is improved gradually through the mutual learning framework, both in aspect-sentiment control and review quality, like the use of 'But' and 'also' which makes the review more coherent.

Dataset	Model	Aspect Accuracy	Sentiment Accuracy
Yelp	ASRG-C_0	0.621	0.610
	ASRG-C_1	0.646	0.635
RateBeer	ASRG-C_0	0.679	0.659
	ASRG-C_1	0.715	0.673

Table 7: Effect of reconstruction reward.

Conclusion and Future work

In this paper, we propose a semi-supervised aspect-level sentiment-controllable review generation method which can utilize both limited labeled data and large-scale unlabeled data to realize aspect-level sentiment control in review generation. In the mutual learning framework, our model enables the generator and classifier enhance each other through the confidence mechanism and reconstruction reward. Experiments on real-world datasets demonstrate that our model can generate aspect-level sentiment-controllable reviews without losing quality.

For future works, we will explore the effectiveness of other generation and classification models applied to our mutual learning framework, since the framework is general and has no requirements for the generator and classifier. Besides, more efforts could be made to expand the aspect-level sentiment labeled data to further improve the performance.

Acknowledgments

This work is supported by the National Key R&D Program of China (2020AAA0105200), Beijing Academy of Artificial Intelligence (BAAI). Huimin Chen is supported by 2019 Tencent Rhino-Bird Elite Training Program.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Bowman, S.; 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.

Cagan, T.; Frank, S. L.; and Tsarfaty, R. 2017. Data-Driven Broad-Coverage Grammars for Opinionated Natural Language Generation (ONLG). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1331–1341.

Chen, H.; Yi, X.; Sun, M.; Li, W.; Yang, C.; and Guo, Z. 2019. Sentiment-controllable Chinese poetry generation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 4925–4931. AAAI Press.

Chen, Y.; and Xie, J. 2008. Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science* 54(3): 477–491.

Cheng, Y.; Xu, W.; He, Z.; He, W.; Wu, H.; Sun, M.; and Liu, Y. 2016. Semi-Supervised Learning for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1965–1974. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1185. URL https://www.aclweb.org/anthology/P16-1185.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.

Dong, L.; Huang, S.; Wei, F.; Lapata, M.; Zhou, M.; and Xu, K. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 623–632.

Gu, J.; Hassan, H.; Devlin, J.; and Li, V. O. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 344–354. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1032. URL https://www.aclweb.org/anthology/N18-1032.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.-Y.; and Ma, W.-Y. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, 820–828.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1587–1596. JMLR. org.

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

Lavie, A.; and Agarwal, A. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 228–231. Association for Computational Linguistics.

Li, J.; Zhao, W. X.; Wen, J.-R.; and Song, Y. 2019. Generating Long and Informative Reviews with Aspect-Aware Coarse-to-Fine Decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1969–1979.

Li, P.; and Tuzhilin, A. 2019. Towards Controllable and Personalized Review 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)*, 3228–3236.

Li, Y.; Li, C.; Zhang, Y.; Li, X.; Zheng, G.; Carin, L.; and Gao, J. 2020. Complementary Auxiliary Classifiers for Label-Conditional Text Generation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lipton, Z. C.; Vikram, S.; and McAuley, J. 2015. Generative concatenative nets jointly learn to write and classify reviews. *arXiv preprint arXiv:1511.03683* .

Luo, F.; Li, P.; Zhou, J.; Yang, P.; Chang, B.; Sun, X.; and Sui, Z. 2019. A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5116–5122. AAAI Press.

Ma, S.; Yang, P.; Liu, T.; Li, P.; Zhou, J.; and Sun, X. 2019. Key Fact as Pivot: A Two-Stage Model for Low Resource Table-to-Text Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2047–2057. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1197. URL https://www.aclweb.org/anthology/P19-1197.

McAuley, J.; and Leskovec, J. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, 165–172.

McAuley, J.; Leskovec, J.; and Jurafsky, D. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, 1020–1025. IEEE.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. 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)*, 188–197.

Ni, J.; Lipton, Z. C.; Vikram, S.; and McAuley, J. 2017. Estimating reactions and recommending products with generative models of reviews. In *Proceedings of the Eighth International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), 783–791.

Ni, J.; and McAuley, J. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 706–711.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Qader, R.; Portet, F.; and Labbé, C. 2019. Semi-Supervised Neural Text Generation by Joint Learning of Natural Language Generation and Natural Language Understanding Models. In *Proceedings of the 12th International Conference on Natural Language Generation*, 552–562.

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.

Tang, J.; Yang, Y.; Carton, S.; Zhang, M.; and Mei, Q. 2016. Context-aware natural language generation with recurrent neural networks. *arXiv preprint arXiv:1611.09900* .

Tilk, O.; and Alumäe, T. 2017. Low-Resource Neural Headline Generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, 20–26.

Tran, V.-K.; and Nguyen, L.-M. 2018. Dual Latent Variable Model for Low-Resource Natural Language Generation in Dialogue Systems. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 21–30. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/K18-1003. URL `https://www.aclweb.org/anthology/K18-1003`.

Wang, K.; and Wan, X. 2018. SentiGAN: generating sentimental texts via mixture adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4446–4452.

Wang, Z.; and Zhang, Y. 2017. Opinion Recommendation Using A Neural Model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1626–1637.

Zang, H.; and Wan, X. 2017. Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation*, 168–177.

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.