One Comment from One Perspective: An Effective Strategy for Enhancing Automatic Music Comment

Tengfei Huo[†]*, Zhiqiang Liu[‡], Jinchao Zhang[‡] and Jie Zhou[‡]
[†]CAS Key Lab of Network Data Science and Technology
[†]Institute of Computing Technology, Chinese Academy of Sciences
[†]University of Chinese Academy of Sciences, Beijing, China
[‡]Pattern Recognition Center, WeChat AI, Tencent Inc, China
huotengfei19s@ict.ac.cn
{zhiqliu,dayerzhang,withtomzhou}@tencent.com

Abstract

The automatic generation of music comments is of great significance for increasing the popularity of music and the music platform's activity. In human music comments, there exists high distinction and diverse perspectives for the same song. In other words, for a song, different comments stem from different musical perspectives. However, to date, this characteristic has not been considered well in research on automatic comment generation. The existing methods tend to generate common and meaningless comments. In this paper, we propose an effective multi-perspective strategy to enhance the diversity of the generated comments. The experiment results on two music comment datasets show that our proposed model can effectively generate a series of diverse music comments based on different perspectives, which outperforms state-of-the-art baselines by a substantial margin.[1]

1 Introduction

In recent years, neural networks have achieved great success in natural language generation (NLG), which can be applied in many real-world scenarios, such as poetry generation, dialogue generation, comment generation, and so on. Music comment generation is a sub-task of NLG. High-quality comments can effectively increase the popularity of music and the activity of the platform (Zeng et al., 2019).

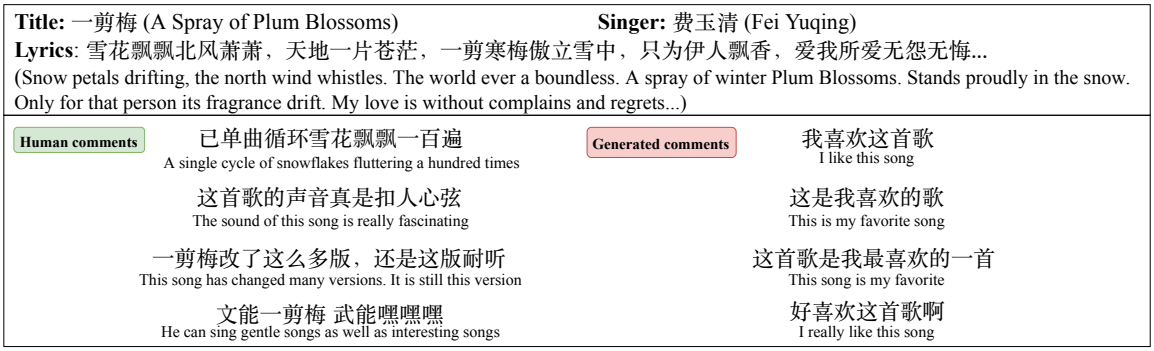


Figure 1: Examples of human comments and generated comments.

However, there is a gap between human comments and generated comments based on the existing model. The music comment generation task's existing models tend to generate general but meaningless comments, such as "It is nice". As an example, an excerpt from human music comments and automatically generated comments based on the seq2seq model is given in Figure 1. Obviously, these human comments on the left are more attractive and diverse, benefiting from the creation with multiple musical perspectives: the listener's behavior, song melody, version of music, and style of lyrics. In human music

*This work was done when Tengfei Huo was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
[1]The datasets and code are available at https://github.com/htfhxx/CommentPerspective

Proceedings of the 28th International Conference on Computational Linguistics, pages 2889–2899
Barcelona, Spain (Online), December 8-13, 2020

comments, there exists excellent distinction and diverse perspectives for the same music content. There goes a saying that *there are a thousand Hamlets in a thousand people's eyes*. As a result, the music comment generation is typically regarded as a one-to-many generation task.

Nowadays, researchers have noticed such problems and tried to solve them via multiple methods in dialogue generation. Some of them have utilized topics, keywords, meta-words, and other information during the generation process to improve performance (Xing et al., 2017; Mou et al., 2016; Xu et al., 2019). Some researchers try to optimize the decoding process (Vijayakumar et al., 2016; Li et al., 2016b) or reorder the candidate sequences after the decoding (Yao et al., 2016; Song et al., 2017). However, these approaches cannot significantly improve the model's performance on diversity. Some methods model one-to-many relationships by multiple latent variables and conform to the dialogue generation scenario (Zhou et al., 2017; Zhou et al., 2018; Chen et al., 2019). They tried to add multiple latent mechanisms or multi-mapping mechanisms between the encoder and decoder of the seq2seq architecture. These one-to-many mapping modules can capture a variety of similar generation modes to a certain extent in dialogue generation. Nevertheless, there is so much overlap of aspects between the text generated through these mapping modules, such as topics and description methods, which is more intolerable in comment generation than in dialogue generation.

This paper aims to bridge the gap between human and machine comments via a multi-perspective mechanism. We define the different language styles, views or aspects of the human comment creation as the musical perspectives, such as *emotional perspectives*, *content theme perspectives*, *lyrics style perspectives*, etc. Besides, compared with other scenarios, the distinction between various comments needs to be more significant in this task for the same input content.

In detail, to better simulate human behaviors and generate more diverse comments, we construct an effective multi-perspective mechanism. The proposed model consists of a music sequence information encoder with a multi-perspective extraction mechanism and a decoder to generate different comments. There is a significant difference between the training stage and the inference stage in our model. In the training stage, the model extracts the perspective that is more conform to the music content and selects the perspective suitable for optimization through the posterior information of the target comment. Besides, we also design a distinction loss function between the perspective extraction components. Our model maximizes the difference between the extraction components by minimizing the loss so that each component exerts a different effect. In the inference stage, the model can separately generate different comments based on different perspectives that are optimized in the previous training stage. Our proposed model not only fits the situation of generating music comments but also dramatically reduces the duplication and redundancy between perspective components. Finally, it can simulate multiple perspectives and generate diverse comments.

Overall, the contributions of this paper are listed as follows:

• As far as we can see, we are the first to improve music comment generation's diversity through a multi-perspective mechanism.

• We propose a novel comment generation model based on multiple-perspective extraction training. The proposed model improves the quality of music comments and makes different perspective modules generate diverse comments.

• The automatic evaluation results show that our model is better than baselines on both datasets. Further analysis and manual evaluation show that the difference between automatically generated comments has been indeed improved significantly.

2 Proposed Method

2.1 Model Overview

First, we define the task of generating music comment. Given an input sequence containing music information $X = \{x_1, x_2, ..., x_T\}$, which contains the song title, author, and lyrics, we hope that our model can generate the corresponding music comment $Y = \{y_1, y_2, ..., y_{T'}\}$. x_i and y_j for $i = 1, 2, ..., T$, $j = 1, 2, ..., T'$are words. T and T' are the lengths of the input sequence and output sequence.

We aim to generate a series of comments Y from multiple perspectives given the textual sequence

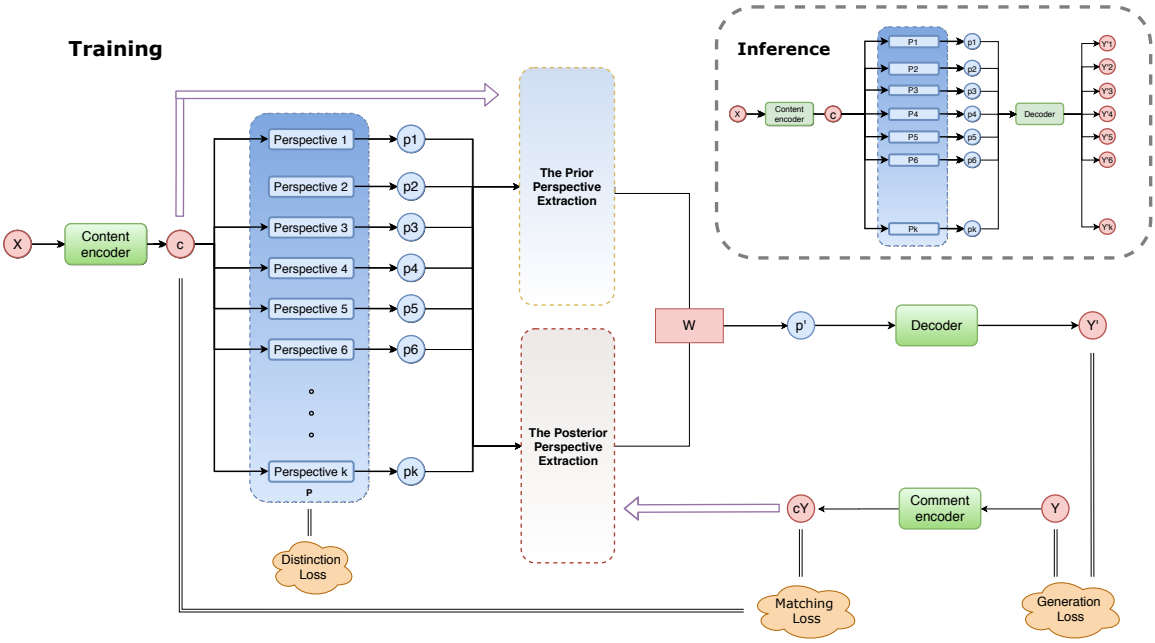


Figure 2: Illustration of our model

X. Figure 2 illustrates the architecture of our model. In the training stage, the input sequence X is encoded and converted to multi-perspective vectors through a multi-perspective mechanism, which simulates commenting for the music based on the concerned aspects and language styles. These multi-perspective vectors need to decode after two ways of extraction. The first way extracts the part of semantic vectors related to the input sequence. The second extracts the part through the posterior information of the target comment. These extracted vectors are finally used in the decoder to generate vivid and diverse comments.

Besides, to avoid duplication and redundancy of these perspective components, we optimize the multi-perspective mechanism by a Distinction Loss Function. The final loss function consists of three parts: Generation Loss, Distinction Loss, and Matching Loss. Among them, Generation Loss is the decoder's negative log-likelihood loss function, and Matching Loss is an auxiliary loss function to project music content and music comment into the same perspective vector space.

In the inference stage, the input sequence X is encoded and directly converted into multi-perspective vectors. All of the multi-perspective vectors individually generate different comments. In this stage, there are no two ways of perspective extraction. We generate corresponding comments for each perspective component. In this way, our model can simultaneously generate comments from multiple perspectives.

2.2 Encoders

The proposed model includes a music content encoder and a comment encoder. The comment encoder is only used in the training stage. Both of them use a single-layer bidirectional GRU (Cho et al., 2014), and the learning parameters are not shared. For the music content encoder, the i-th hidden states of forward and backward GRU are computed by:

$$\overrightarrow{h}_i = GRU_{forward}\left(\overrightarrow{h}_{i-1}, e\left(x_i\right)\right) \tag{1}$$

$$\overleftarrow{h}_i = GRU_{backward}\left(\overleftarrow{h}_{i-1}, e\left(x_{T-i+1}\right)\right) \tag{2}$$

where $e\left(x_i\right) \in \mathcal{R}^d$ is the embedding of word x_i and d is the dimension of embeddings. Then corresponding hidden states of forward and backward GRU are concatenated as the i-th hidden states h_i. Finally we use $x = h_T = [\overrightarrow{h}_T, \overleftarrow{h}_1]$ as the semantic representation of input sequence. Samely, the semantic representation of music comment Y is y.

2.3 Multi-Perspective Mechanism

We use multiple linear functions to implement the multi-perspective mechanism. We call these linear functions the perspective components. We update the parameters of these perspective components with the help of two ways of perspective extraction. The simple linear function can also simulate a variety of comment perspectives to generate diverse comments.

Each linear function performs a linear transformation on the semantic representation of music content x, thereby capturing the underlying regularities in semantics:

$$p_k = W_k x + b_k \tag{3}$$

where $W_k \in R^{l_p \times T}$, and $b_k \in R^{l_p}$ are the parameters of the k-th perspective component. l_p and T is dimension of p and x. And $\{p_i\}_{i=1}^{K}$ are the multi-perspective vectors.

2.4 Perspective Extraction

These different perspective vectors contain potential information in language expression methods or aspects of description. Eventually, they are fed into the decoder to generate vivid and diverse comments. However, if these unprocessed vectors are directly used in the decoding process, there will be no difference between these perspective components. Therefore, perspective extraction needs to be performed, and appropriate multi-perspective vectors are extracted for parameter updates in the training stage.

These multi-perspective vectors go through two ways of extraction. In the prior perspective extraction, considering that not all the potential information is suitable for all songs, the model learns to extract a part of suitable vectors for each input sequence. For example, for a sad piece of music, we cannot use a comic style to comment. In the posterior perspective extraction, we utilize the posterior perspective extraction mechanism similar to Chen et al. (2019) to pick out suitable perspective vectors for parameter updates.

The Prior Perspective Extraction. The prior perspective extraction selects the perspective vectors related to the input sequence. So we need to model $\mathcal{P}_\alpha(p_i|x)$ and find an appropriate weight for each perspective vector:

$$\mathcal{P}_\alpha\left(p_i|x\right) = \frac{\exp g_\alpha\left(p_i, x\right)}{\sum_{j=1}^{K} \exp g_\alpha\left(p_j, x\right)} \tag{4}$$

$g_\alpha\left(p_i, x\right)$ is used to get the similarity between p_i and x. Inspired by Zhou et al. (2017), in order to avoid over-fitting of $g_\alpha\left(p_i, x\right)$, we add two learnable parameter matrix and use the Maxout activation function (Goodfellow et al., 2013):

$$g_\alpha\left(p_i, x\right) = p_i^{\mathrm{T}} W_z z \tag{5}$$

$$z = \left[\max\left\{\widetilde{z}_{2j-1}, \widetilde{z}_{2j}\right\}\right]_{j=1,2,\cdots,T} \tag{6}$$

$$\tilde{z} = W_x x + b_x \tag{7}$$

where $W_z \in R^{l_p \times T}$, $W_x \in R^{2T \times T}$, $b_x \in R^{2T}$. l_p and T are dimensions of p and x.

The correlation of the encoded music content x and multi-perspective vectors can be obtained through the above process. Moreover, the weights of each perspective vector can be calculated. Thereby it achieves the purpose of the prior perspective extraction.

The Posterior Perspective Extraction. The posterior perspective extraction is similar to Chen et al. (2019). We use the posterior information of the target comment to extract the multi-perspective vectors. We calculate the correlation between the target comment and each perspective vector. Then we use softmax normalization to get the probability of extracting the perspective vector:

$$\mathcal{P}_\beta\left(p_i|y\right) = \frac{\exp g_\beta\left(p_k, y\right)}{\sum_{i=1}^{K} \exp g_\beta\left(p_i, y\right)} \tag{8}$$

where g_β is dot product operation.

Perspective Extraction. The two-way perspective extraction can give different weights to different perspective vectors. Besides, we fuse the extracted result of the two ways to obtain the final perspective vector:

$$p' = p_z, \quad z = \underset{i=1,2,..,k}{\arg\max}(\tau_i) \tag{9}$$

$$\tau_k = softmax(W_\alpha \mathcal{P}_\alpha + W_\beta \mathcal{P}_\beta + b) \tag{10}$$

where $W_\alpha, W_\beta \in R^{l_p \times l_p}$ and $b \in R^{l_p}$ are both learnable parameters. l_p is dimension of p. We extract the perspective vector with the highest probability and fed it into the decoder. For the backpropagation of the sampling process, we use Gumbel-Softmax reparametrization (Jang et al., 2017) to obtain the probability. So that different samples can reasonably optimize various perspective components in the training stage.

2.5 Decoder

The comment decoder uses a unidirectional GRU:

$$s_j = GRU\left(s_{j-1}, e\left(y_{j-1}\right), c_j\right), \quad s_0 = p' \tag{11}$$

where s_j is the hidden state of GRU, c_j is the context vector of time step j. We use the extracted multi-perspective vector p' as the initial state of the hidden layer. The generation probability of each time step in the decoding process is:

$$\mathcal{P}\left(y_j | y_{0:j-1}, X, P\right) = softmax\left(s_j, c_j\right) \tag{12}$$

where $y_{0:j-1}$ is generated text before the time step.

However, in the inference stage, there is no longer a perspective extraction process. Each perspective vector will be fed into the decoder and generate a comment. Through multi-perspective comments generated by multiple perspective components, we can enhance the performance of automatic music comments.

2.6 Distinction Loss

To further increase each perspective component's difference, we add a regularization term about the multi-perspective mechanism for the loss function. We aim to enhance the difference between the rows of the parameters matrix of linear functions.

Inspired by Lin et al. (2017), we perform a dot product operation on the parameters matrix and its transpose. Then the result of the dot product minus the identity matrix. We add the final result as a penalty term for enlarging the difference between rows:

$$\mathcal{L}_D = \left\| \left(DD^T - I \right) \right\|_F \tag{13}$$

$$D_i = \frac{\exp(p_i)}{\sum_{j=1}^{K} \exp(p_j)} \tag{14}$$

$\| \cdot \|_F$ stands for the Frobenius norm of a matrix.

2.7 Overall Loss Function

In posterior perspective extraction, the encoded target comment y is required. So we need to project the music comment y into the same perspective vector space, and we use the Matching Loss (Chen et al., 2019). For each input sequence, we randomly collected some negative samples, which aims to guide the input content and current target comment mapping to the same semantic space. Matching Loss is the negative log-likelihood of relevance for the encoded music content and encoded comments:

$$\mathcal{L}_M = -\log \mathcal{P}(r = 1 | X, Y) + \log \mathcal{P}\left(r = 1 | X, Y^-\right) \tag{15}$$

$$\mathcal{P}(r = 1 | X, Y) = \sigma(x \cdot y) \tag{16}$$

In the decoding process, the loss function of generating comments is:

$$\mathcal{L}_G = -\log \mathcal{P}\left(Y|X, P\right) \tag{17}$$

Therefore, the total loss function of our model is:

$$\mathcal{L} = \mathcal{L}_G + \mathcal{L}_M + \gamma \mathcal{L}_D \tag{18}$$

3 Experiments

3.1 Datasets and Setups

We construct two Chinese music comment datasets for all experiments. One is the QQ Music comment dataset, while the other is the NetEase Cloud Music comment dataset. In detail, we collect the QQ music comment dataset, including about 61,618 pieces of song information-comment data from the online music website[2] and collect about 205,085 comments from the NetEase Cloud music website[3].

The details of the datasets are shown in Table 1. The two datasets have a significant difference, proving that our model can show good results on the different music comment datasets.

Dataset	QQ Music	NetEase Cloud Music
# Total	61,618	205,085
# Train	49,295	164,069
# Dev	6,162	20,508
# Test	6,161	20,508
Average Length of Comments	38.96	13.41

Table 1: Statistics of datasets. # is the number of samples.

Meanwhile, in our experiments, the size of word embedding is 200, and we initialize the word embedding from Tencent AI Lab Embedding Corpus[4]. The number of perspective components is 20. The hidden size is set to 1024. We use Adam optimizer (Kingma and Ba, 2014), and the learning rate is set to 0.0002. We set a dropout rate of 0.3 and use the Beam Search to generate all samples, and the beam size is set to 10. The coefficient γ of Distinction Loss is 0.001 in the NetEase Cloud Music dataset and 0.00005 in the QQ Music dataset. The model generally reaches the optimality of the validation set within ten epochs. Additionally, we choose the better one between the model of the 10-th epoch and the model with the lowest loss on the verification set.

3.2 Baselines

For the experimental comparisons, we compare our model with the following baselines:

• **Seq2Seq** (Qin et al., 2018): This model follows the framework of the sequence-to-sequence model with attention.

• **CVAE** (Zhao et al., 2017): The conditional variational auto-encoder based approach.

• **VMED** (Le et al., 2018): This model associates each memory read with a mode in the latent mixture distribution at each timestep. It can capture the variability observed in sequential data.

• **MMPMS** (Chen et al., 2019): A state-of-art multi-mapping mechanism model. It focuses on selecting the corresponding mapping module by the target response. Following their setting, we set the number of mapping modules to 20.

3.3 Evaluation Metrics

We use two kinds of evaluation methods: automatic evaluation and manual evaluation. For automatic evaluation, we used BLEU-1/2 (Chen and Cherry, 2014) to test the percentage of overlap of unigram and

[2]https://y.qq.com/

[3]https://music.163.com/

[4]https://ai.tencent.com/ailab/nlp/en/index.html

bigram between the generated comment and ground truth. We also use Dist-1/2 (Li et al., 2016a) to test the richness of unigram and bigram in all the comments generated. For manual evaluation, inspired by Liu et al. (2019) and Zhou et al. (2018), we adopt the following four manual evaluation metrics:

• **Fluency**: Whether the comments are fluent and whether there are severe grammatical errors.

• **Coherence**: Whether the generated comments conform to the scenario of music. How relevant the comment is to the music content.

• **Meaningfulness**: Whether the generated comments have rich meaning and detailed content.

• **Distinction**: Whether there exist significant differences between the generated comments for the same music input content. The greater the difference, the higher the score.

All the above metrics are scored on a five-point scale, and we take the average of scores as the final result. We construct a manual test set containing 250 generated comments for each model, which belongs to 50 input samples. We invite five human experts to provide scores according to the above criteria, and the average score for each metric is computed.

4 Results

4.1 Evaluation Results

	QQ Music				NetEase Cloud Music			
	BLEU-1	BLEU-2	Dist-1	Dist-2	BLEU-1	BLEU-2	Dist-1	Dist-2
Seq2Seq	0.155	0.060	0.0001	0.002	0.103	0.066	0.010	0.079
CVAE	0.178	0.062	**0.009**	0.066	0.127	0.058	0.012	0.145
VMED	0.187	0.065	0.008	0.054	0.130	0.061	0.013	0.152
MMPMS	0.204	0.146	0.005	0.029	0.157	0.126	**0.017**	0.275
OURS	**0.409**	**0.344**	0.008	**0.081**	**0.198**	**0.191**	0.016	**0.290**

Table 2: Results of automatic evaluation

Table 2 shows the automatic evaluation results of the two datasets. It can be easily observed that our proposed model obtains a higher BLEU score and Dist-2 score than baselines on the two music comment datasets. The value of Dist-1 is also close to the highest value of baselines. The improvement of the BLEU score reflects that our model can generate more informative comments, which may be attributed to the design of prior perspective extraction. Moreover, the Dist score shows that the generated comments are diverse, and the vocabulary is rich enough.

	QQ Music				NetEase Cloud Music			
	F	C	M	D	F	C	M	D
Seq2Seq	3.92	2.24	2.08	1.00	4.31	3.11	2.62	1.82
CVAE	**3.98**	**2.80**	2.60	1.04	4.01	3.32	2.84	1.34
VMED	3.93	2.29	2.85	1.02	4.06	3.34	2.74	1.20
MMPMS	3.87	2.31	2.36	3.06	4.06	3.24	2.65	3.40
OURS	3.91	2.48	**2.99**	**3.25**	**4.41**	**3.40**	**2.90**	**3.86**

Table 3: Result of manual evaluation. **F** means *Fluency*. **C** stands for *Coherence*. **M** represents *Meaningfulness* while **D** represents *Distinction*.

The human evaluation results, as shown in Table 3, indicate that our model has better performance on manual evaluation. In terms of Fluency and Relevance of the QQ music dataset, our model score is low. According to the observation of the results, we find that it is because the comment needs to generate is too long, and some repeated text fragments appear frequently. The metric of Distinction has a significant improvement compared to baselines, which shows that the comments our model generates are very diverse.

4.2 Further Analysis

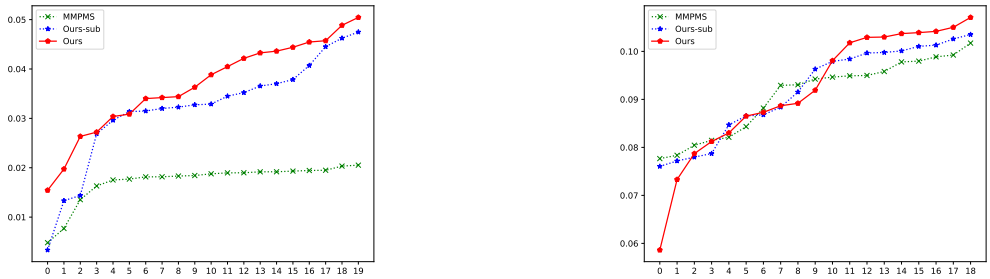
To validate that the difference of perspective components in our model has been significantly improved and our model can generate multi-perspective comments, we further analyze the results generated by the models. Two similar models are selected to compare with our model. One is the baseline MMPMS, and the other is a sub-model that removes Distinction Loss from the proposed model.

	NetEase Cloud Music	QQ Music
MMPMS	0.0903	0.0673
OURS-sub	0.0881	0.0642
OURS	**0.0794**	**0.0625**

Table 4: Average Similarity of comments generated by the different perspectives.

In order to measure the difference of comments based on different perspectives, we proposed a new metric called **Average Similarity** that is defined as the average similarity between multiple comments generated based on different perspective components for the same music input content.

We randomly select 200 samples. For each sample, we calculate the average similarity of generated comments. In detail, we use BERT (Devlin et al., 2018) to vectorize the comments and obtain 768-dimensional vectors. Then similarity is measured by the method of cosine similarity. The result in Table 4 shows that the average of our model is lower than other models, which indicates that multiple comments generated by the same case in our proposed model are more diverse. The difference in perspective components has significantly improved.



(a) QQ Music dataset (0-19)

(b) NetEase Cloud dataset (1-19)

Figure 3: Dist-2 of all the perspective components.

In addition to comparing the multiple comments corresponding to different perspectives, we can also analyze the overall difference of results set corresponding to every perspective component.

We compare the metric Dist-2 of results set corresponding to every perspective component. In order to facilitate comparison, we have sorted them. As can be seen from the figure 3, the difference between perspective components from the proposed model is noticeable. Figure (b) only shows nineteen dots because the first value from MMPMS is too slow to be an abnormal value. The polyline of our model is more tending to a straight line of $y = kx+b$, which means there is a uniform degree of difference between each component. Furthermore, the effect on the long text from the QQ Music dataset is more significant. It is worth noting that our improvements expand the difference between perspective components and significantly improve the diversity of the text generated from each perspective component.

4.3 Case Study

Table 4 presents generated comments from Seq2seq, MMPMS and our model. We select five comments for each model. The five comments generated by the Seq2seq model come from the top-5 in the process of beam searching. Moreover, we select the top 5 of the smoothest sentences from the results of the single input for MMPMS and our model. It can be seen that the comments generated by Seq2Seq, and MMPMS inevitably focus on the same perspective, and even generate duplicate phrases and words. For

Title: 微笑的力量 (The power of smile) Singer: 陈谦文 (Chen Qianwen)

Lyrics: 晴时多云偶阵雨，偶尔失去太阳的勇气，就算挫折浸湿了翅膀，梦想也不曾停止远扬，难免会哭泣，像倾盆大雨淋着雨，我陪你前行，仰望着雨后彩虹...
Cloudy occasional showers, occasionally losing the sun's courage, even if the frustration soaked the wings, the dream never stopped flying, it will inevitably cry, like a downpour rain, I accompany you, looking forward to the rainbow after the rain.

Seq2seq	MMPMS	Our Model
你让我拥有微笑的力量 You give me the power to **smile**	好温柔的声音 So **gentle** voice	你的微笑的力量很棒 The **power** of your **smile** is great
我们一起飞到世界的尽头 We **fly** to the end of the world **together**	整个宇宙的声音真的超级温暖 The sound of the entire universe is really **warm**	我觉得很好听啊 I think it **sounds** good
我喜欢你的声音 I **like** your voice	每天把梦全都照亮 **Light up** all **dreams** every day	希望有雨后彩虹陪你前行 I hope **rainbow rain** will accompany you after rain
你让我拥有微笑的力量。 You give me the power to **smile**.	梦想也不散 **Dreams** are not gone	飞进了灿烂星空把梦全都照亮 Fly into the **starry sky** and **light up** all the dreams
我们一起飞 We **fly together**	星空把梦全都照亮 Starry sky **illuminates** all **dreams**	我不羡慕你少年 Young man, I don't **envy you**

Figure 4: Examples of comments generated by different models for the same music content

instance, all the 3th-5th comments from the MMPMS model describe about "梦想(dream)" and "照亮(light up)". However, the comments generated by our model are meaningful and diverse. Meanwhile, our model can also generate comments from more perspectives or topics.

5 Related work

The text generation based on the Seq2Seq model tends to create general text. For example, existing models on open-domain comment generation always produce repetitive and uninteresting comments (Lin et al., 2019). Li et al. (2019) model the input news as a topic interaction graph and generate comments with a graph-to-sequence model. Lin et al. (2019) retrieve informative and relevant comments by leveraging user-generated data. However, many researchers try to model a one-to-many relationship to solve this similar problem in dialogue generation tasks. In detail, Xing et al. (2017) use topics to simulate prior human knowledge and guide them to form informative responses. In contrast, Mou et al. (2016) utilize pointwise mutual information to extract words as keywords and decode the response based on the keywords. Liu et al. (2018) propose a neural knowledge diffusion model to introduce knowledge into dialogue generation. Zhang et al. (2018) apply an explicit specificity control variable into a seq2seq model to generate responses at different specificity levels. Besides, Xu et al. (2019) enhance the seq2seq architecture with a goal tracking memory network to incorporate meta-words into generation. The above methods aim to enhance the diversity of generated results via adding specific structures, which has already achieved a particular improvement.

In recent years, some researchers try to construct multiple latent mechanisms to model the one-to-many relationship and generate diverse results. Among them, Tao et al. (2018) propose a novel Multi-Head Attention Mechanism (MHAM), which aims at capturing multiple semantic aspects from the user utterance. Zhou et al. (2017) develop an encoder-diverter-decoder framework, which is used to encode the input into mechanism-aware context, and decode the responses with the controlled styles or topics. Based on the previous work, Zhou et al. (2018) add the filter modules and obtain better results, which selects a subset from all mechanisms to make it contain enough mechanisms to generate multiple style responses. Chen et al. (2019) try to get accurate optimization of latent mechanisms and design a kind of mapping selection method for a multi-mapping mechanism.

6 Conclusion

This paper proposes an effective multi-perspective strategy to enhance automatic music comment and achieve that one comment is from one perspective. The strategy solves the problem of generating common but meaningless comments in the automatic music comment to some extent. We reform the one-to-many modeling mechanism and make it fit the situation of generating music comment.

In conclusion, our model bridges the gap between human and machine comments via the multi-

perspective mechanism, simulating various perspectives to generate diverse music comments. Experiment results show that our method has achieved excellent performance in both automatic evaluation and manual evaluation. The proposed model can also be applied to other text generation scenarios such as news comment generation and poetry generation.

References

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4918–4924. ijcai.org.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C. Courville, and Yoshua Bengio. 2013. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1319–1327. JMLR.org.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hung Le, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2018. Variational memory encoder-decoder. In *Advances in Neural Information Processing Systems*, pages 1508–1518.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Wei Li, Jingjing Xu, Yancheng He, Shengli Yan, Yunfang Wu, and Xu Sun. 2019. Coherent comments generation for chinese articles with a graph-to-sequence model. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4843–4852. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2019. Learning comment generation by leveraging user-generated data. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7225–7229. IEEE.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498.

Zhiqiang Liu, Zuohui Fu, Jie Cao, Gerard de Melo, Yik-Cheung Tam, Cheng Niu, and Jie Zhou. 2019. Rhetorically controlled encoder-decoder for modern chinese poetry generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1992–2001.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3349–3358. ACL.

Lianhui Qin, Lemao Liu, Victoria Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic article commenting: the task and dataset. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 151–156. Association for Computational Linguistics.

Yiping Song, Zhiliang Tian, Dongyan Zhao, Ming Zhang, and Rui Yan. 2017. Diversifying neural conversation model with maximal marginal relevance. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 169–174.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5416–5426. Association for Computational Linguistics.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong. 2016. An attentional neural conversation model with improved specificity. *arXiv preprint arXiv:1606.01292*.

Wenhuan Zeng, Abulikemu Abuduweili, Lei Li, and Pengcheng Yang. 2019. Automatic generation of personalized comment based on user profile. In Fernando Emilio Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 229–235. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1108–1117.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ganbin Zhou, Ping Luo, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. 2018. Elastic responding machine for dialog generation with dynamically mechanism selecting. In *Thirty-Second AAAI Conference on Artificial Intelligence*.