

# Learning to Generate Diverse and Authentic Reviews via an Encoder-Decoder Model with Transformer and GRU

Kaifu Jin, Xi Zhang

*Key Laboratory of Trustworthy Distributed Computing and Service,  
Ministry of Education, Beijing University of Posts and Telecommunications  
Beijing, China  
{jkf, zhangx}@bupt.edu.cn*

Jiayuan Zhang

*Wellcome Centre for Human Genetics  
The University of Oxford  
Oxford, UK  
jiayuan.zhang@stx.ox.ac.uk*

**Abstract**—Fake reviews automatically generated by machine learning models can be manipulated to influence the customers opinions, which is a great threat to online review platforms like social networks and E-commerce websites. Previous review generation methods generally adopt either businesses information (e.g. location and products) or existing review texts from consumers as inputs, while currently no approach that utilizes both types of information has been reported. As business information can help generated reviews gain relevance, and existing user reviews help improve the diversity of generated reviews, we envision that an integration of these two types of information is likely to result in a better review generator. To this end, we propose an encoder-decoder model to produce authentic and diverse reviews, which applies Transformer and mutative Gated Recurrent Unit (GRU) to encode the business information and the customer reviews, respectively. In addition, to address the lack of suitable metrics for evaluating the diversity of reviews, we developed a novel text diversity metric called DMet. Our experiments on Yelp dataset demonstrate that the model we developed can produce reviews with better quality and diversity as compared to existing methods, and DMet is able to closely match human judgment in evaluating text diversity.

**Index Terms**—Fake Review, Text Generation, Opinion Spam, Natural Language Generation

## I. INTRODUCTION

Fake review is one of the major threats to online review systems. Consumers usually prefer products or services with high rating scores on online crowd-sourced review forums or E-commerce like Yelp, Amazon, and eBay. These online platforms contain a large number of reviews written by consumers who had purchased certain products or services, in order to influence the decisions of consumers who are not familiar with the product or service. Driven by the importance of online reviews, fake reviews have emerged. Early fake reviews were written by crowdturfers who were paid to give positive or negative comments to certain businesses to mislead customer perception of the business [30]. With the development of text generation techniques, review-generating algorithms have replaced expensive crowdturfers. As a result, automatic

generation and detection of fake reviews have become a hot research topic.

Currently, there are three major challenges regarding review generation: (1) **Relevance**. Reviews should be closely related to the subject (e.g. products or businesses). For example, a review describing sushi and sashimi will be inappropriate when commenting on a Mexican restaurant. (2) **Diversity**. Generated reviews should be different from each other and from existing user reviews, requiring some level of novelty. (3) **Quality**. The vocabulary and grammar of generated reviews should be correct, with a legible style to disguise itself as authentic user reviews.

Some of existing studies only use user review texts to generate fake reviews [1] [4], while others make use of the context information, such as the review rating, restaurant name, city, state and food tags [2]. It lacks of research works that combine both of them. The context information can help to produce relevant reviews, while the review texts are useful to produce diverse reviews. Therefore, it is beneficial to integrate context information together with the existing reviews to obtain both relevant and diverse reviews.

Motivated by the above idea, we develop an encoder-decoder model [11] that encodes both the context information and the reviews. In particular, the context information is encoded with transformer encoder blocks [9] and the reviews are encoded by mutative GRU [35]. Both of the encodes work together to produce a comprehensive hidden embedding, which serves as an input to the decoder component. This model is capable of generating relevant and diverse reviews.

One of the problems we face in developing the review generating model is the lack of a proper metric for the evaluation of review diversity. Previous works have used Dist [34] and BLEU-recall [21] [4] to measure text diversity, but they are not comprehensive enough to evaluate the diversity of generated reviews. Neural models [36] [37] have also been proposed to measure the diversity, but they require a large manually scored dataset for training purpose, which limit their

applicability.

To realize an accurate diversity measurement of generated reviews, we design a metric for text diversity evaluation named DMet. DMet takes into account three major factors regarding the diversity of the review texts, including inter-sequence diversity, intra-sequence diversity, and sentence length, in order to produce a comprehensive diversity score to then given machine-generated text.

In our experiments, DMet demonstrates extraordinary accuracy in terms of the Pearson and Spearman correlations with manually annotated diversity scores, which outperforms its competitors. As reported by human annotation and DMet, the reviews generated by our model are more deceptive, with higher relevance, quality, and diversity than previous results.

Our contributions are listed as follows:

- We design an encoder-decoder review-generation model combining the Transformer model and the GRU encoder to encode the features from both user reviews and business contexts.
- We propose a metric named DMet for measuring the diversity and novelty of machine-generated text.
- We conduct extensive experiments on the Yelp dataset, and evaluate the performance of our review-generation model and DMet. Experimental results suggest that our model is able to produce more diverse and more authentic reviews than previous works do, and DMet is a better approximation to the human judgment of diversity than existing diversity metrics.

## II. RELATED WORK

### A. Fake Review Generation

Review generation is an important part of natural language generation and information content security. Some researchers [19] generate peer reviews for scientific papers to provide references for journal editors and reviewers. Some researchers [23] [24] generate code reviews to help developers understand programs when reading source code. Moreover, based on the E2ENLG dataset<sup>1</sup> [29], many researchers [25] [26] [27] [28] generate descriptions, which can also be considered as a type of review, from dialog act-based meaning representations of restaurants. Radford et al. [31] develop a method to discover sentiment of generated reviews from Amazon product review dataset [18] and Ni et al. [6] generate fake review from the same dataset by means of considering the relationship between user and item and the effect between user-aspect and item-aspect.

Our generation of fake review is based on the Yelp dataset. Verified on the same dataset, DP-GAN [4] has been proved to be capable of generating long and diverse reviews by combining GAN (Generative Adversary Network) [32] and reinforcement learning. In addition, [1] use a 2-layer fLSTM [10] to generate prototype reviews and customize them by replacing nouns related to the topic word (e.g. “food”) in reviews with corresponding nouns in existing reviews of the target

business. A hyper-parameter in softmax called temperature is adjusted to generate reviews of different degrees of novelty.

Juuti et al. [2] propose a method called NMT-Fake based on open-NMT-py, an open-source neural machine translation framework. Unlike earlier research, they make full use of the businesses information which they call context, including names, tags, and locations of businesses. The NMT-model is a Seq2Seq model [12] containing an RNN encoder and an RNN decoder. For one given review, Juuti et al. use its score and the context of the business of the review as the input of NMT-model, and the review itself as the target. Besides, they creatively propose multiple penalties as hyper-parameters in the process of generating reviews. By adjusting these penalties, safer or newer reviews can be generated. Furthermore, the experiments of Juuti et al. prove that their model has higher quality generated reviews than model of Yao et al. [1]

Our approach differs from these methods mainly in two aspects. Firstly our model is based on the Transformer, a model originally proposed to perform neural machine translation tasks [13], which is more capable of handling longer text sequences than traditional RNN [16] and LSTM [15] [17]. Secondly, we set two separate encoders to process different sources of data, i.e. transformer encoder process the context and GRU the review text. The results from our experiments suggest that our model is able to generate reviews of high quality and diversity.

### B. Diversity Metrics

The evaluation of the text quality is an integrative part of studying automatically generated text. Traditional evaluation methods include BLEU [5], ROUGE [3] and METEOR [38], which are all belong to the family of Word Overlap-based Metrics. These methods are mainly based on the degree of overlap of the words between generated text and reference text. There are also some Embedding-based Metrics, like Greedy Matching, Embedding Average and Vector Extrema [8]. These methods calculate the similarity between generated text and reference text, making them suitable for tasks like machine translation and abstract generation. However, these metrics are shown to correlate poorly with human judgments in reviews generation or dialogue generation. Therefore, we need some indicators to measure the diversity and novelty of generated text to ensure diversified text can be generated.

For the evaluation of diversity and novelty, many studies [7] [14] [39] [40] [45] [48] use Dist (Distinct), a metric of the proportion of distinct n-grams. In addition to the Dist, Zhang et al. [7] propose and use Entropy as a measure of diversity, which reflects how evenly the empirical n-gram distribution is for a given sentence. The Entropy is also used by later researchers [39] [46] [47] [48]. Zhang et al. [20] improve Diverse Text Generation by Self Labeling Conditional Variational Auto Encoder. Their evaluation metric of diversity is based on Dist and BLEU-recall. However, as Dist, Entropy, and BLEU-recall do not take into account intra-sequence repeats and sentence length, we believe they are unable to

<sup>1</sup><http://www.macs.hw.ac.uk/InteractionLab/E2E/>

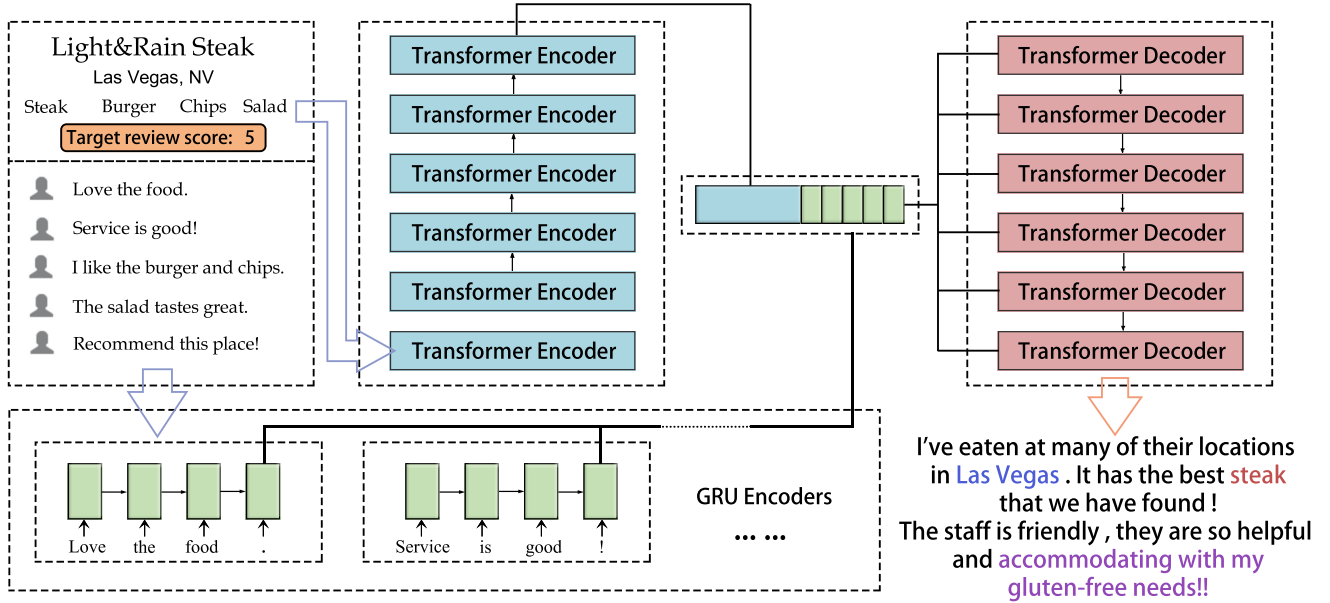


Fig. 1. Architecture of our model. The Light&Rain Steak is an example of a restaurant display on Yelp. Its context is the input of the Transformer Encoders and the reviews are the inputs of GRU Encoders. Then the outputs of the Transformer Encoders and the hidden states of the GRU Encoders are concatenated and fed to the Transformer Decoders, whose final output is the generated review. The review contains words captured in the context and comprises deceptive description (e.g. “accommodating with my gluten-free needs!!”). Details like the input for transformer Decoder and multi-head self-attention are not depicted in this figure.

give a comprehensive judgment on the diversity and novelty of generated text.

Based on the idea of Word Overlap-based Metric, our evaluation metric measures the diversity and novelty of generated texts by calculating word overlap, intra-sequence similarity penalty, and sentence length correction. Also, our evaluation metric is model-independent, enabling convenient migration to other tasks like review generation and dialog generation.

### III. MODEL

#### A. Dataset

We use the dataset provided by Yelp Dataset Challenge<sup>2</sup> Round 13. Yelp Dataset is a well-known dataset used for review generation tasks. This dataset contains information about businesses, reviews, users, images, etc. on yelp.com. About 188k businesses and 6M reviews are included in the dataset. We select business data within the “restaurant” category along with the corresponding reviews for pre-processing.

The pre-processing of the input dataset consists of two components: pre-processing of contexts and pre-processing of reviews. Our pre-processing method on the contexts is mainly based on Juuti’s method [2]. That is, for each restaurant and target review score (e.g. 5), we extract up to five existing reviews corresponding to the target score. Then these reviews are combined with the restaurant information (e.g. Light&Rain Steak Las Vegas NV Steak Burger Chip Salad) to serve as an input of the model. For the ease of handling, we use two tags to separate the context. <SEP1> is used to separate the context

from the input reviews, and <SEP2> is used to separate the individual reviews in the input reviews. Finally, the example of pre-processing consequence is as below:

```
5 Light&Rain Steak LasVegas NV Steak Burge Chip Salad<SEP1>
Love the food.<SEP2>Service is good!
<SEP2>I like the burger and chips.<SEP2>The salad tastes great.
<SEP2>Recommend this place!
```

The data after pre-processing contains about 3.65 million pairs of inputs and targets similar to the example above.

#### B. Problem Definition

We aim to generate the reviews corresponding to a specific rating score for a specific restaurant, by giving the restaurant information and several reviews under the restaurant. Formally, we describe the review generation task as follows. Given a group of inputs  $\{c_1, c_2, \dots, c_i, \dots, c_M\}$  which consist of contexts and target scores, and a group of review lists  $\{R_1, R_2, \dots, R_i, \dots, R_M\}$ , where  $R_i$  is a list of up to five reviews, our goal is to generate reviews  $\{r_1, r_2, \dots, r_i, \dots, r_M\}$ .

#### C. Framework

The model framework we use is based on the Seq2Seq model. Compared to previous models that use RNN or GAN for review generation, the input of the Seq2Seq model is not necessarily equal in length to the target, making it possible to increase the relevance of the generated reviews to the business by adding context to the input.

In specific, we consider the context as the primary input for generating the review text, and for reference, some of the

<sup>2</sup><https://www.yelp.com/dataset/challenge>

reviews are added in the input. Therefore, the Transformer Encoder is selected as the encoder responsible for processing the context, while the mutative GRU Encoder is responsible for processing the input review information. Therefore, the majority of hidden information provided to the Decoder comes from the Transformer Encoder, and the rest comes from the GRU Encoder. We use Transformer Decoder as the decoder in our model as it can capture longer text features than traditional RNN, which we believe will result in the longer and more reliable text to be generated. Our model architecture is shown in Figure 1.

#### D. Transformer Encoder and GRU Encoder

Our encoder comprises of two parts to handle different types of input data. The first part is the 6-layer Transformer encoder, whose input is the pre-processed context and target score. The Transformer encoder will generate input embedding and position encoding for each tag of the context, and then create new representations of the context by self-attention and multi-head attention. For every given input  $c_i$ , the output of the Transformer encoder is  $t$  and the multi-head self-attention mechanism is  $attn_i$ , which are calculated as:

$$t_i, attn_i = TransformerEncoder(c_i), \quad (1)$$

where  $t_i \in \mathbb{R}^{S_i \times H_t}$ ,  $S_i$  stands for sequence length of the input and  $H_t$  is the hidden size of the Transformer encoder.

The second part is the GRU encoder, consisting of up to 5 sub-encoders numbered in the order of 1-5, which is considered as an index number. Each sub-encoder processes a review with the matching index number in the input review list  $R_i$ . If the index number exceeds the number of reviews in the input review list, the corresponding sub-encoder will be unavailable for this input. Every available sub-encoder will return a hidden state  $h_i^j$ , which is calculated as:

$$h_i^j = GRUSubEncoder_j(R_i^j), \quad (2)$$

where  $j$  is the index number of sub-encoder and review which ranges from 1 to 5.  $GRUSubEncoder_j$  represents the  $j$ -th sub-encoder,  $h_i^j \in \mathbb{R}^{l \times H_g}$ ,  $l$  is the layer number of GRU and  $H_g$  is the dimension of hidden states of every sub-encoder. Here every encoder is a one-layer GRU, so  $l = 1$ .

#### E. Transformer Decoder

The decoder of our model is a 6-layer Transformer decoder and the input of the Transformer decoder is the review text  $r_i$ . Similar to other common decoders in Encoder-Decoder models, our Transformer decoder is also used to predict the target words. The input of the Transformer decoder is the sentence embedding and position embedding of a specific review, which is relevant to the context, the score and the corresponding review list. We concatenate the  $t_i$  and every  $h_i^j$  as one of the inputs of the Multi-head attention, where the other two are the  $attn_i$  and processed input of the Transformer decoder.

The Transformer decoder input one word and generate one word at a time from left to right just like the RNN. For a given input review text  $r_i$  of the Transformer decoder, we denote the input word of the  $s$ th step as  $r_i^s$ , and the output word as  $\hat{r}_i^{s+1}$ . The calculation process in Transformer decoder can be simplified as:

$$\hat{r}_i^{s+1} = TransformerDecoder(d_i, h_i^j; r_i^s) \quad (3)$$

where the  $d_i$  is calculated as:

$$d_i = t_i + \sum r_i^j \quad (4)$$

In the generation stage, we input the start token as the first word to Transformer decoder, and then we input the output word of the Transformer decoder at the next step. Finally, the review text will be generated step by step.

### IV. DMET

#### A. Requirements of Diversity Metric

We propose a diversity metric named DMet for measuring the diversity of generated reviews. To ensure generated reviews are diversified, three aspects should be noted. First, generated reviews need to be different from user reviews under the same subject. Second, the closer the length of a fake review is to that of real reviews, the more deceptive it will be. Third, the generated review may include internal repeats, so the intra-sequence similarity should be measured.

Our metric can also be applied to other text generator tasks that require diversity scoring, such as dialog generation. There are three corresponding requirements for the generation diversity in a wider range of text generation tasks. We refer to a generated text as a “candidate” and some references texts of the candidate as “references”. Accordingly, the three requirements will be:

- 1) The candidate should be distinct from the references (i.e. the inter-sequence diversity).
- 2) To ensure the candidate is informative, the length of the candidate should be similar or above the typical length of references (i.e. the appropriate length of candidate).
- 3) Local repetition within the sentence should be avoided in the candidate, especially for long-length candidates (i.e. the intra-sequence diversity).

A comparison of the key factors taken into account by various diversity metrics is shown in Table II. The Dist includes considerations of intra-sequence novelty and inter-sequence novelty, but not the positive effect of appropriate candidate length on diversity. BLEU-recall takes into account inter-sequence diversity, but not the other two factors. Entropy is a good indication of diversity in terms of intra-sequence diversity and proper length but ignores the diversity between sentences.

#### B. Details of DMet

The diversity evaluation metric we design is largely inspired by BLEU. Our evaluation metric is based on the calculation of the weighted word-overlap, to which three diversity indicators are added, including weighted word-overlap count,

TABLE I  
HYPER-PARAMETERS SETTINGS

Word Dimension	GRU Hidden Size	Transformer Hidden Size	Batch Size	Learning Rate	Optimizer	Residual Dropout
256	512	256	32	0.01	Adam	0.3

TABLE II  
COMPARISON OF VARIOUS DIVERSITY METRICS

Metrics	intra-sequence	inter-sequence	Appropriate Length
Dist	✓	✓	✗
BLEU-recall	✗	✓	✗
Entropy	✓	✗	✓
DMet	✓	✓	✓

intra-sequence penalty, and sentence length correction. We first calculate proportion the weighted  $n$ -grams overlap in all  $n$ -grams as a modified precision score  $P_n$  for all candidate sentences:

$$Count_{gram}(n-gram) = f_\lambda(\Delta) \times Count_{cand}(n-gram), \quad (5)$$

$$P_n = \frac{\sum_{C \in candidates} \sum_{n-gram \in C} Count_{gram}(n-gram)}{\sum_{C^* \in candidates} \sum_{n-gram^* \in C^*} Count(n-gram^*)}, \quad (6)$$

$f_\lambda$  is an attenuation function with parameter  $\lambda$ , and  $n$ -gram means one  $n$ -word span in a sentence. The symbol  $\Delta$  represents the total count of a given gram in all references:

$$\Delta = \sum_{r \in references} Count_{ref}(n-gram), \quad (7)$$

$$f_\lambda(\Delta) = \lambda^\Delta, \quad (8)$$

where  $\lambda \in (0, 1)$ .

According to the above formula,  $f_\lambda(\Delta)$  can be regarded as a weight related to the count of the given  $n$ -gram in references. The more frequently the  $n$ -gram appears in references, the smaller  $f_\lambda(\Delta)$  will be, corresponding to a smaller value for the numerator in equation (4). The parameter  $\lambda$  controls the rate at which  $f_\lambda(\Delta)$  decreases as  $\Delta$  increases. Here we make  $\lambda = 0.5$ .

Then we calculate the geometric mean of precision score from uni-gram to 4-gram:

$$P_{avg} = \left( \prod_{n=1}^N P_n^{w_i} \right)^{\frac{1}{\sum_{i=1}^N w_i}}, \quad (9)$$

where  $N = 4$  and the  $w_i$  means the weight of different precision scores. In addition, to avoid meaningless intra-

sequence repeats within each candidate, we add the intra-sequence Penalty (IP) to punish this situation:

$$IP = \frac{\sum_{C \in candidates} g(split(C))}{\sum_{C^* \in candidates}}, \quad (10)$$

where the  $split(\bullet)$  stands for the process to split the candidate  $C$  to some shorter sentences. The penalty function  $g(\bullet)$  we use is ROUGE-2 [3]. Besides, for generated sentences, we believe that longer, non-repetitive sentences are generally more informative, so we compute the Length Correction (LC) of each candidate:

$$LC = \frac{\sum_{C \in candidates} \Phi(C)}{\sum_{C^* \in candidates}}, \quad (11)$$

where  $\Phi(C)$  is the length correction of a candidate:

$$\Phi(C) = \left( \frac{l_C}{l_{ref}^{mean}} \right)^{\frac{l_{ref}^{min}}{l_{ref}^{mean}}}, \quad (12)$$

where  $l_C$  is the length of a candidate, while  $l_{ref}^{mean}$  and  $l_{ref}^{min}$  represent the average and minimum length values of corresponding references, respectively.

Finally, we simply multiply  $P_{avg}$ ,  $IP$ , and  $LC$  to obtain the diversity metric DMet:

$$DMet = P_{avg} \times IP \times LC \quad (13)$$

In the DMet,  $P_n$ , IP, and LC considered the inter-sequence diversity, the intra-sequence diversity and the appropriate length respectively.

## V. EXPERIMENT

### A. Training Details

To implement our model and baselines, we are grateful to use the following open-source toolkits in Python in the process of text processing and review generation: PyTorch<sup>3</sup>, OpenNMT<sup>4</sup>, SpaCy<sup>5</sup>, Scipy<sup>6</sup>, Matplotlib<sup>7</sup>, and Pandas<sup>8</sup>.

Hyper-parameter settings are shown in Table I. Each of our Transformer encoder and decoder is a stack of 6 Transformer layers, and the dimension of encoder output is 512. Every GRU sub-encoder is a 2-layer GRU with a hidden size of 512. We also add Dropout [33] before and after the Transformer

<sup>3</sup><https://github.com/pytorch/pytorch>

<sup>4</sup><https://github.com/OpenNMT/OpenNMT-py>

<sup>5</sup><https://github.com/explosion/spaCy>

<sup>6</sup><https://github.com/scipy/scipy>

<sup>7</sup><https://github.com/matplotlib/matplotlib>

<sup>8</sup><https://github.com/pandas-dev/pandash>

layer to avoid over-fitting. The dropout rate and the residual dropout rate are both set to 0.3. Besides, we use reviews as part of inputs in the training stage but not the generating stage for the cold-start purpose. On the hardware side, the CPU and GPU we use are Intel Xeon E5-2620 and NVIDIA GTX 1080, respectively.

### B. Human Annotation

In order to accurately evaluate the authenticity and diversity of generated reviews, and to measure if DMet closely resembles human judgment, we invited some volunteers to label the reviews. Review-labeling is divided into two tasks: identifying fake reviews from real ones and diversity scoring of machine-generated reviews.

1) *Task-1*: The first task asks volunteers to judge the authenticity of each review from a list containing both real and generated reviews. We prepared 345 contexts and 2,648 real reviews of businesses. Then we use the contexts as model inputs to get 345 fake reviews from NMT-Fake and same number reviews from our model. Next, we randomly shuffle the real reviews and the machine-generated reviews under each context, and then we ask the volunteers to browse each context and mark out associated fake reviews. A detailed description of the questionnaire we used along with representative screenshots is included in Appendix: Review Annotation Task Design.

In order to simulate the perspective of real customers, the proportion of machine-generated reviews is not revealed, as is the case in actual review platforms. The experimental result is analyzed in Model Evaluation.

2) *Task-2*: The second task is to evaluate the diversity of machine-generated reviews. We label machine-generated reviews that appeared in Task-1 and ask volunteers to score them based on their diversity. Five grades of increasing diversity are used: 2, 4, 6, 8 and 10, with 2 representing no diversity and 10 representing great diversity.

### C. Experiment on Diversity Metrics

We calculate the correlation coefficient between DMet and human annotation from task-2, which reflects the similarity between DMet and human judgment to some extent. The correlation coefficient between each of Dist-1, Dist-2, Entropy-4 and BLEU-recall<sup>9</sup>, and the human evaluation result, are also calculated for comparison purpose. Since Dist-1 and Dist-2 have little meaning on one single sentence, we randomly select 10 machine-generated reviews to calculate Dist-1 and Dist-2 and calculate the average annotated scores of these 10 reviews as the human score. This random selection is repeated 300 times. The final results are shown in Table III and Figure 2.

According to the experimental results, DMet has the strongest correlation with human metrics for diversity, in terms of both the Pearson correlation coefficient and Spearman correlation coefficient. Note that the correlation coefficients

<sup>9</sup>Dist-1 and Dist-2 represent the Dist of uni-gram and bi-gram. Entropy-4 represents the Entropy of 4-gram. All the setting are according to the references.

TABLE III  
CORRELATION BETWEEN AUTOMATIC METRICS AND HUMAN ANNOTATION

Metrics	Pearson (p-value)	Spearman (p-value)
Human Score	0.8068 (<0.01)	0.7866 (<0.01)
Dist-1	-0.1828 (<0.01)	-0.1778 (<0.01)
Dist-2	0.0142 (0.7216)	0.0310 (0.4376)
BLEU-recall	0.5427 (<0.01)	0.5708 (<0.01)
Entropy	0.5892 (<0.01)	0.5841 (<0.01)
DMet	<b>0.6236</b> (<0.01)	<b>0.6311</b> (<0.01)

TABLE IV  
AUTOMATIC EVALUATION RESULTS OF DIFFERENT MODELS

Metrics	Our model	NMT-Fake
Dist-1	0.0499	<b>0.0538</b>
Dist-2	<b>0.3272</b>	0.3167
Uniq-1	<b>2088</b>	1578
Uniq-2	<b>13377</b>	8993
Average Length	<b>44.45</b>	31.17
BLEU-recall	<b>1.193</b>	1.1488
Entropy-4	<b>3.6406</b>	3.2788
DMet	<b>0.5903</b>	0.5293

between the Dist-1 and human judgment are negative and close to zero. Apart from the reason mentioned previously [7], we believe this is due to the fact that as the text gets longer, high-frequency words are more likely to appear than low-frequency words. This leads to the phenomenon that Dist-1 will decrease in response to the increase in the average length of the generated text. However, in the task of review generation, it is generally believed that longer, less repetitive reviews have higher diversity, so we argue that Dist-1 is not suitable for diversity evaluation in the review generation task.

### D. Model Evaluation Results

We compare our model with the state-of-the-art model proposed by Juuti et al. which is called NMT-Fake. The NMT-Fake utilize the context to generate fake reviews stayed on-topic. The comparison is from the aspect of automatic evaluation and the aspect human evaluation.

1) *Automatic Evaluation*: Previously in Experiment on Diversity Metrics, we have demonstrated that DMet has good performance in measuring diversity. Now in this subsection, we collect 941 contexts containing target review score and business information as the input and obtained 941 generated reviews from our model and from NMT-Fake. These generated reviews are then subjected to evaluation with DMet and other diversity metrics. We include no reviews as input during the stage of generating reviews, unlike the training stage, in order to evaluate the generated reviews under a uniform cold-started condition. In addition to the aforementioned diversity indicators, we also use Uniq-1 and Uniq-2, representing the number of unique uni-grams and bi-grams, which are the numerators of Dist-1 and Dist-2.

The experimental results are shown in Table IV. Our model outperforms NMT-Fake on DMet, BLEU-recall, Dist-2, while unsurprisingly, our score on Dist-1 is lower than NMT-Fake. Due to reasons mentioned in Experiment on Diversity Metrics,

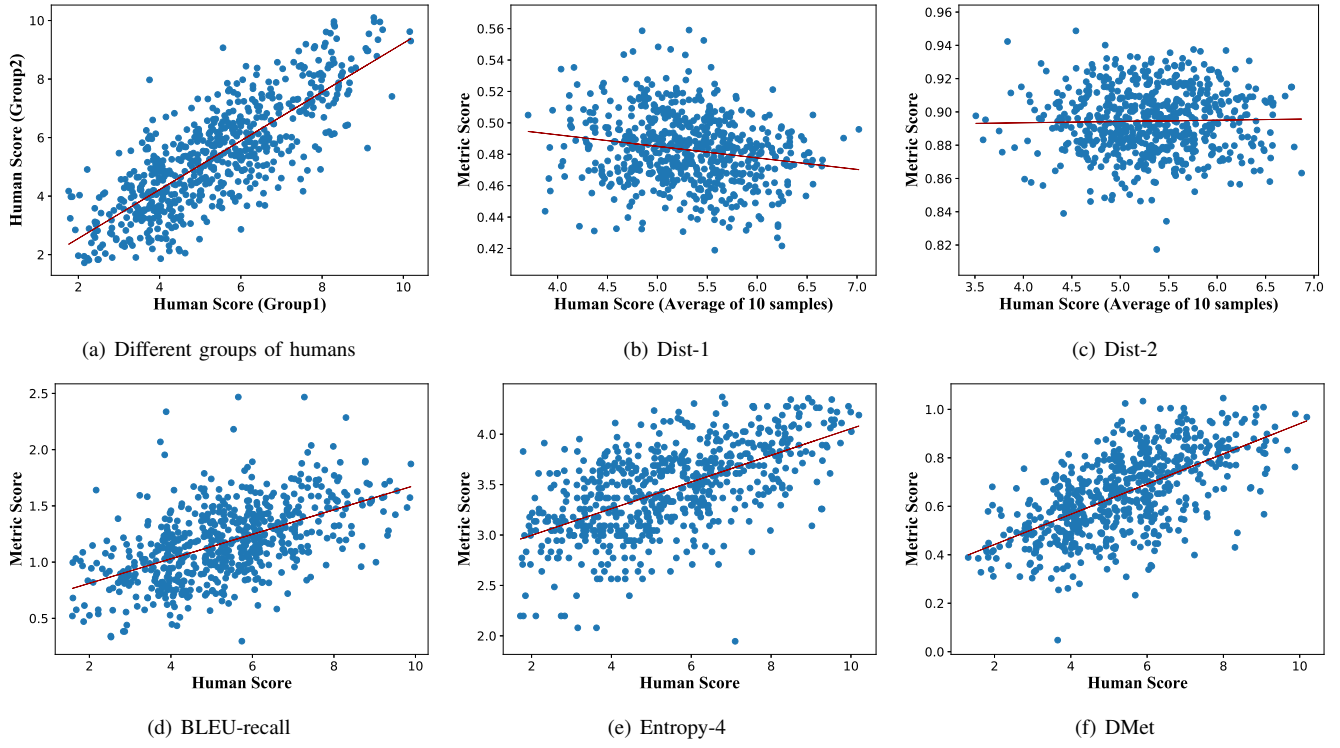


Fig. 2. The correlation coefficients between diversity metrics and human annotations. (a) Volunteers have been randomly divided into two groups and the correlation coefficient is calculated between the two. (b)(e) Scatter plots of existing metrics against average human scores. (f) Scatter plots of proposed diversity metric DMet against average human scores. Inspired by [36], a Gaussian noise of  $N(0, 0.3^2)$  is added for better visualization of point density. For (a) (d) (e) (f), each point is associated with a machine-generated review. For (b) (c), each point is associated with 10 randomly selected machine-generated reviews.

although the unique uni-gram(Uniq-1) of the text generated by our model is 32.32% higher than the unique uni-gram of the text generated by NMT-Fake, the longer (42.61%) length of text generated by our model actually lowers its Dist-1 score.

2) *Human Evaluation*: Human judgment serves as an important reference in evaluating the quality of the machine-generated text. We use the human annotation results obtained in task 1 to evaluate if generated reviews can closely resemble authentic reviews. We designate machine-generated reviews as positive classes and real reviews as negative classes, and then we calculate the precision, recall, and F1-score. Precision denotes the proportion of machine-generated reviews in all reviews that are considered as fake, while recall stands for the probability that a machine-generated review is detected as a fake review. The F-1 score is the harmonic average of the precision and recall, representing the overall accuracy of the judgement. So for precision, recall, and F1-score, the lower the better.

According to the experimental results, the recall of our model is significantly lower than that of NMT-Fake ( $p = 0.0024$ , Pearson's chi-squared test). Therefore, compared to those generated by NMT-Fake, it is considerably more difficult to discriminate reviews generated by our model from real reviews.

In addition, we have statistically summarized the diversity annotation of machine-generated reviews by all volunteers in task 2. The results are shown in Table V and Figure 3. As

TABLE V  
HUMAN EVALUATION RESULTS OF DIFFERENT MODELS

Human Evaluation	Our model	NMT-Fake
Precision	<b>0.1383</b>	0.1785
Recall	<b>0.3275</b>	0.4397
F1-score	<b>0.1945</b>	0.2539
Average Diversity Score	<b>5.4031</b>	4.8889

shown in Table V, the average diversity of reviews generated by our model is higher than the reviews generated by NMT-Fake. As can be seen from Figure 4, significantly higher proportions of high-diversity reviews are generated by our model than with NMT-Fake.

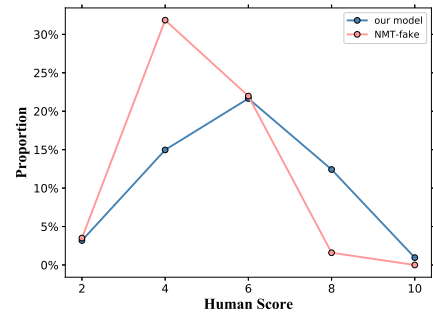


Fig. 3. The distribution of human-annotated diversity scores for reviews generated by our model and NMT-fake

TABLE VI  
EXAMPLES GENERATED BY OUR MODEL AND THE STATE-OF-THE-ART MODEL ON THE YELP DATASET

Model	Score	Fatburger Vaughan ON Burgers Restaurants	Cibo Wine Bar Toronto ON Restaurants Nightlife Italian Wine Bars Mediterranean Bars
NMT-Fake	1	Worst customer service. Took 20 minutes to receive a sandwich, no one even came by for my order or anything.	The staff at this place sucks. They don't care about how to treat customers and the wait time can not be worth it. It is very loud, but you get what they pay for!
	2	The <b>burgers</b> are okay. I had high expectations for this place, and it's just ok... nothing really special either! They do not have any of the other options in the menu to offer but they don't seem very good.	Very disappointing for a place that specializes on food. Service was mediocre, but overall it was good and reasonably priced.
	3	<b>Burgers</b> are great. They have an awesome selection and they've got good service, however it can get quite busy at times so I would recommend getting there early or expect to wait.	I'm a fan of the decor and ambience. The service is always good, however their food quality is just okay and overpriced compared to other locations. It's not worth the money for sure!
	4	Food is good but not the greatest service. I've had better <b>burger</b> at other locations and they were pretty much all their toppings. They have many options for meat eaters as well! Will come again to try more menu items.	I'd give it 5 stars, but the food is a bit overpriced.
	5	The food was very good. I had their <b>burger</b> and poutine, it's really nice to have one in the area.	The service was excellent. Food came fast, the food and <b>wine</b> were great! I would definitely come back again. Highly recommend.
Our Model	1	The <b>burger</b> is horrible, I don't know how they cook it. It tastes funny but not worth your money or calories for a single meal. They also charge extra to add salt and pepper on top! Not going back again!!	The staff is rude. I ordered pizza at 11:30 and they said it was not even busy, but when she told us we were waiting for our food that wasn't worth it or anything like this.
	2	I went on a Sunday afternoon, ordered 2 large fries for lunch and it was very salty ( the <b>burgers</b> were good ). My husband had the double patty with cheese and bacon added to it. The onion rings are great! However there's only one person working in cash register so that should have taken care of customers	I was really excited to try this place out. The atmosphere and decor were nice but the service wasn't great, it just didn't seem like any other restaurant in town where you would find a better <b>Italian</b> food.
	3	The quality of the ingredients were good. The <b>burger</b> was a little on the salty side, but overall I would come again	The atmosphere was nice, but I'm not impressed. We sat at the bar so we went to watch football games on the TV. Our server was friendly enough for our group of 8 people and everyone seemed happy with each other while they waited on us! The wine list could use some improvement though.
	4	Great <b>burgers</b> . The fries were great! I'm sure they have a lot to choose from, which is why I didn't give 5 stars.	Went there with my wife. The place is very busy, but we were seated quickly and served right away in front of our group. We had a nice dinner at the lounge area which made it great for us! Food was delicious and I would recommend to everyone who loves <b>Italian</b> food so this restaurant has been here forever!!
	5	Just love the <b>burgers</b> here! I always get the veggie patty with fries, and a drink. The staff is friendly. They also offer halal food at this place for their customers which are very important when you order to eat in or out of line up your car	We ordered the calamari, bruschetta&pasta. Everything is fresh and delicious! Service has improved since we were there last year.

## VI. CONCLUSION AND FUTURE WORK

In this paper, our work focuses on generating fake reviews automatically. We propose a model to generate diverse and authentic reviews, which is an Encoder-Decoder model containing Transformer and mutative GRU encoders, as well as Transformer decoders. Furthermore, we design DMet to accurately evaluate the diversity of the generated text. Both the review-generation model and DMet display satisfactory performance in our experiments. We use automatic evaluation and human evaluation to demonstrate that our model can generate more diverse and authentic reviews than the current state-of-

the-art method [2], and DMet matches human judgment better than existing diversity metrics.

As the theories and practice of natural language processing are undergoing rapid development, the Transformer and GRU we use are not the most advanced text generation models to date. In the current work, we do not utilize pre-trained language models such as ELMo [43], BERT [41], GPT [42], and XLNet [44], which have driven the progress of many tasks in natural language processing. Hopefully, using an appropriate pre-training language model may help further improve the quality of generated reviews in future works. And



from the perspective of detecting false comments, machine-generated reviews have reached a stage where it is difficult to identify them from authentic reviews with human judgment only, as shown in the above results. Previously some detection strategies have been proposed [1] [2], but their viability is largely limited to situations where the defender can obtain a considerable amount of the known fake reviews generated by the attacker model, which is seldom the case. Therefore, in addition to inspecting the review texts, any abnormal account behaviors should not be missed in fake review detection<sup>10</sup>.

## VII. ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Program of China (No. 2017YFB0803301) and Natural Science Foundation of China (No. 61976026, No. U1836215) and 111 Project (B18008).

## REFERENCES

- [1] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. Automated crowdturfing attacks and defenses in online review systems. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2017.
- [2] Mika Juuti, Bo Sun, Tatsuya Mori, and N. Asokan. Stay on-topic: Generating context-specific fake restaurant reviews. In Proceedings of the 23rd European Symposium on Research in Computer Security (ESORICS), pages 132151, 2018.
- [3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pages 7481, 2004.
- [4] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 39403949, 2018.
- [5] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311318, 2002.
- [6] Jianmo Ni, Julian McAuley, 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), 2018.
- [7] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujuan Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In Advances in Neural Information Processing Systems, pages 18131823, 2018.
- [8] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023, 2016.
- [9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Advances in Neural Information Processing Systems 30, pages 6000-6010. Curran Associates, Inc, 2017.
- [10] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735-1780, 1997.
- [11] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: 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), 2014.
- [12] I. Sutskever, O. Vinyals, and Q. V. Le, Sequence to sequence learning with neural networks, in Advances in neural information processing systems, pp. 3104-3112, 2014.
- [13] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473, 2014.
- [14] Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, Rui Yan. Are Training Samples Correlated? Learning to Generate Dialogue Responses with Multiple References. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 38263835, 2019.
- [15] Gongbo Tang, Mathias Miller, Annette Rios, and Rico Sennrich. Why self-attention? a targeted evaluation of neural machine translation architectures. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.
- [16] Tom Mikolov, Martin Karafit, Luk Burget, Jan ernock, Sanjeev Khudanpur. Recurrent neural network based language model. In 11th Annual Conference of the International Speech Communication Association, pages 1045-1048, 2010.
- [17] Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. An evaluation of neural machine translation models on historical spelling normalization. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1320-1331, 2018.
- [18] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pages 43-52, 2015.
- [19] Bartoli, Alberto, Andrea DeLorenzo, Eric Medvet, and Fabiano Tarlao. Your Paper Had Been Accepted, Rejected or Whatever: Automatic Generation of Scientific Paper Reviews. International Cross Domain Conference and Workshop (CD-ARES), 2016.
- [20] Yuchi Zhang, Yongliang Wang, Liping Zhang, Zhiqiang Zhang, Kun Gai. Improve Diverse Text Generation by Self Labeling Conditional Variational Auto Encoder. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.
- [21] Zhao, T.; Zhao, R.; and Eskenazi, M. 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), volume 1, 654-664, 2017.
- [22] Albert Gatt and Emiel Krahmer. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research (JAIR), 61:65-170, 2018.
- [23] Xing Hu, Ge Li, Xin Xia, David Lo, and Zhi Jin. Deep code comment generation. In Proceedings of the 26th Conference on Program Comprehension, pages 200-210. ACM, 2018.
- [24] Uri Alon, Omer Levy, and Eran Yahav. code2seq: Generating sequences from structured representations of code. In International Conference on Learning Representations, 2019.
- [25] Juraj Juraska, Panagiotis Karagiannis, Kevin K. Bowden, and Marilyn A. Walker. A Deep Ensemble Model with Slot Alignment for Sequence-to-Sequence Natural Language Generation. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 152162, 2018.
- [26] Shang-Yu Su and Yun-Nung Chen. Investigating linguistic pattern ordering in hierarchical natural language generation. In 7th IEEE Workshop on Spoken Language Technology (SLT), 2018.
- [27] Sam Wiseman, Stuart Shieber, and Alexander Rush. Learning neural templates for text generation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3174-3187, 2018.
- [28] Shang-Yu Su, Kai-Ling Lo, Yi-Ting Yeh, and Yun-Nung Chen. Natural language generation by hierarchical decoding with linguistic patterns. In Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.
- [29] Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, 2017.
- [30] Rinta-Kahila, T., Soliman, W.: Understanding crowdturfing: The different ethicallogics behind the clandestine industry of deception. In Proceedings of the 25th European Conference on Information Systems, 2017.
- [31] Radford, A., Jozefowicz, R., and Sutskever, I. Learning to Generate Reviews and Discovering Sentiment. arXiv preprint arXiv:1704.01444, 2017.

<sup>10</sup>Researchers focusing on the detection of fake reviews are welcome to contact us for the source code of our model.

- [32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems 27, 2014.
- [33] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1):1929-1958, 2014.
- [34] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Proceedings of NAACL-HLT 2016, pages 110119, 2016.
- [35] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
- [36] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [37] Xiaowei Tong, Zhenxin Fu, Mingyue Shang, Dongyan Zhao, and Rui Yan. One “ruler for all languages: Multi-lingual dialogue evaluation with adversarial multi-task learning. In Proceedings of the 27th International Joint Conference on Artificial Intelligence pages 4432-4438, 2018.
- [38] Satandeep Banerjee and Alon Lavie. 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, 2005.
- [39] Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, Huan Sun. Reinforced Dynamic Reasoning for Conversational Question Generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 21142124, 2019.
- [40] Zhiliang Tian, Wei Bi, Xiaopeng Li, Nevin L. Zhang. Learning to Abstract for Memory-augmented Conversational Response Generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 38163825, 2019.
- [41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [42] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. URL [https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf)
- [43] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.
- [44] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237.
- [45] Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, Xu Sun. Enhancing Topic-to-Essay Generation with External Commonsense Knowledge. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 20022012, 2019.
- [46] Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyrill Truskovskiy, Alexander Tselousov, Thomas Wolf. Large-Scale Transfer Learning for Natural Language Generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 60536058, 2019.
- [47] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, Jianfeng Gao. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 54275436, 2019.
- [48] Daphne Ippolito, Reno Kriz, Maria Kustikova, Joao Sedoc, Chris Callison-Burch. Comparison of Diverse Decoding Methods from Conditional Language Models. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 37523762, 2019.

## VIII. APPENDIX: REVIEW ANNOTATION TASK DESIGN

The review annotation task is completed via an online questionnaire<sup>11</sup> where we ask volunteers to mark out reviews they believe are machine-generated. The context information of restaurants is provided in the form of sequences of strings. Depending on the specific number of available user reviews, 9 to 45 reviews are presented for annotation under each restaurant.

An example of the first page of the user study is shown in Figure 4 and Figure 5.

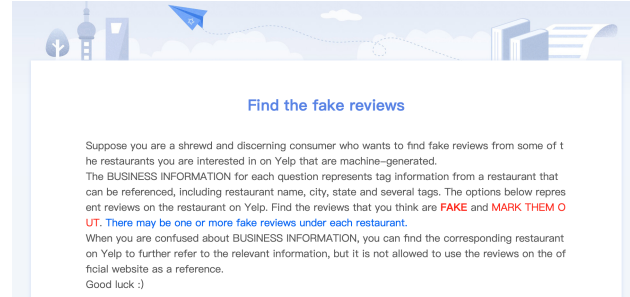


Fig. 4. Screenshots A of the survey. The Task-1 description in human annotations.

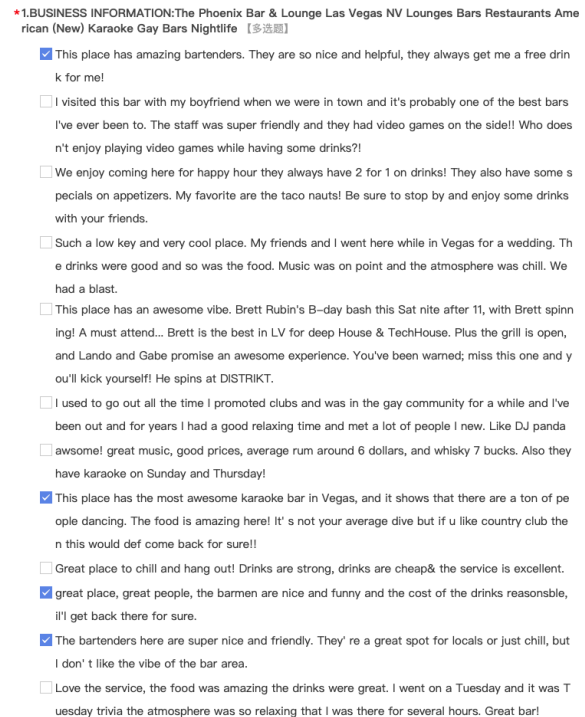


Fig. 5. Screenshots B of the survey. An example of the Task-1 in human annotations. Options marked are machine-generated reviews. Of the four machine-generated comments, the first and third are from NMT-Fake, and the second and fourth are from our model.

<sup>11</sup>The questionnaire tool we use is from <https://www.wjx.cn/>