TopicRefine: Joint Topic Prediction and Dialogue Response Generation for Multi-turn End-to-End Dialogue System

Hongru Wang[1,2,*], **Mingyu Cui**[1,*], **Zimo Zhou**[1], **Kam-Fai Wong**[1,2]
[1]The Chinese University of Hong Kong, Hong Kong, China
[2]MoE Key Laboratory of High Confidence Software Technologies, China
{hrwang,kfwong}@se.cuhk.edu.hk

Abstract

A multi-turn dialogue always follows a specific topic thread, and topic shift at the discourse level occurs naturally as the conversation progresses, necessitating the model's ability to capture different topics and generate topic-aware responses. Previous research has either predicted the topic first and then generated the relevant response, or simply applied the attention mechanism to all topics, ignoring the joint distribution of the topic prediction and response generation models and resulting in uncontrollable and unrelated responses. In this paper, we propose a joint framework with a topic refinement mechanism to learn these two tasks simultaneously. Specifically, we design a three-pass iteration mechanism to generate a coarse response first, then predict corresponding topics, and finally generate a refined response conditioned on predicted topics. Moreover, we utilize GPT2DoubleHeads and BERT for the topic prediction task respectively, aiming to investigate the effects of joint learning and the understanding ability of the GPT model. Experimental results demonstrate that our proposed framework achieves new state-of-the-art performance at the response generation task and the great potential understanding capability of the GPT model.

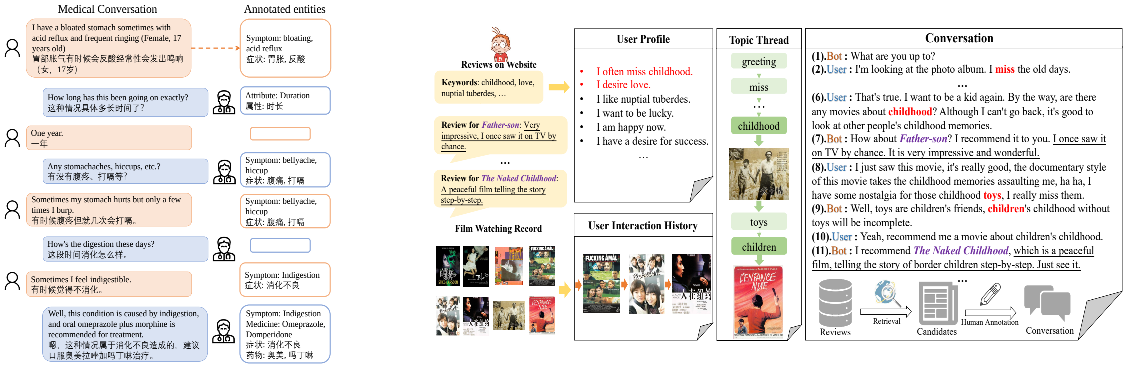
1 Introduction

Natural Language Generation (NLG), is the task of generating language that is coherent and understandable to humans, and has been applied to many downstream tasks such as text summary (Zhang et al., 2019a; Bar-Haim et al., 2020; Cho et al., 2020; Huang et al., 2020; Gholipour Ghalandari and Ifrim, 2020), machine translation (Li et al., 2020; Baziotis et al., 2020; Cheng et al., 2020; Zou et al., 2020), and dialogue response generation (Radford et al., 2019; Zhou et al., 2018b; Tuan et al., 2019; Zhao et al., 2020; Liu et al., 2020a; Wolf et al., 2019).

Recent works in dialogue response generation usually formulate this task as a sequence-to-sequence problem, leading to inconsistent, uncontrollable, and repetitive responses (Ram et al., 2018). Furthermore, each dialogue has its specific goal and each utterance of the dialogue may contain multiple topics, regardless it is an open-domain dialogue or task-oriented dialogue. As shown in *left* part of Figure 1, the patient seeks medical advice from a doctor and informs him of the attributes and symptoms of the specific disease which form the topics of the conversation. Also, some open-domain dialogue systems have specific goals, such as recommendations, education, etc. For example, a conversational agent interacts with a user to recommend some interesting movies (as shown in *right* part of Figure 1). The entire content flow is guided by the topic thread. These various conversational scenarios propose more challenges for the current multi-turn end-to-end dialogue system, necessitating the model's capability to generate a more informative and topic-related response.

Many researchers propose different methods to guide or control the generation of responses conditioned on specific topics. Some representative works consider incorporating topic information into the sequence-to-sequence framework which applies an attention mechanism to all topics (Xing et al., 2017; Dziri et al., 2019). Other works cast this task as a pipeline system, predict the keywords, then capture the topic, and finally retrieve corresponding response (Tang et al., 2019; Zhou et al., 2020). Another line of work focuses on single-turn topic-aware response generation conditioned on previously given topics (Feng et al., 2018; Yang et al., 2019; Huo et al., 2020). All these methods fall short in two ways. Most of these approaches either heavily rely on the non-autoregressive models like BERT (Devlin et al., 2019) to predict topics or utilize the attention mechanism on all pre-defined topics which do not consider the effect of the histor-

[*]These authors contributed equally to this work



I have a bloated stomach sometimes with acid reflux and frequent ringing (Female, 17 years old)
胃部胀气有时候会反酸经常性会发出鸣响（女，17岁）

Symptom: bloating, acid reflux
症状: 胃胀, 反酸

How long has this been going on exactly?
这种情况具体多长时间了?

Attribute: Duration
属性: 时长

One year.
一年

Any stomachaches, hiccups, etc.?
有没有腹疼、打嗝等?

Symptom: bellyache, hiccup
症状: 腹痛, 打嗝

Sometimes my stomach hurts but only a few times I burp.
有时候腹疼但就几次会打嗝。

Symptom: bellyache, hiccup
症状: 腹痛, 打嗝

How's the digestion these days?
这段时间消化怎么样。

Sometimes I feel indigestible.
有时候觉得不消化。

Symptom: Indigestion
症状: 消化不良

Well, this condition is caused by indigestion, and oral omeprazole plus morphine is recommended for treatment.
嗯，这种情况属于消化不良造成的，建议口服奥美拉唑加吗丁啉治疗。

Symptom: Indigestion
Medicine: Omeprazole, Domperidone
症状: 消化不良
药物: 奥美, 吗丁啉

Keywords: childhood, love, nuptial tuberdes, …

Review for ***Father-son***: Very impressive, I once saw it on TV by chance.

…

Review for ***The Naked Childhood***: A peaceful film telling the story step-by-step.

- I often miss childhood.
- I desire love.
- I like nuptial tuberdes.
- I want to be lucky.
- I am happy now.
- I have a desire for success.

…

(1). Bot **:** What are you up to?
(2). User **:** I'm looking at the photo album. I **miss** the old days.
…
(6). User **:** That's true. I want to be a kid again. By the way, are there any movies about **childhood**? Although I can't go back, it's good to look at other people's childhood memories.
(7). Bot **:** How about ***Father-son***? I recommend it to you. I once saw it on TV by chance. It is very impressive and wonderful.
(8). User **:** I just saw this movie, it's really good, the documentary style of this movie takes the childhood memories assaulting me, ha ha, I have some nostalgia for those childhood **toys**, I really miss them.
(9). Bot **:** Well, toys are children's friends, **children**'s childhood without toys will be incomplete.
(10). User **:** Yeah, recommend me a movie about children's childhood.
(11). Bot **:** I recommend ***The Naked Childhood***, which is a peaceful film, telling the story of border children step-by-step. Just see it.
…

Figure 1: **Left:** MedDG Dataset **Right:** TG-ReDial Dataset. Adapted from (Liu et al., 2020a) and (Liu et al., 2020b) respectively.

ical topic path of multi-turn conversations. Besides that, these works do not model the joint distribution of attribute model $p(a|x)$ and unconditional language model $p(x)$, which is proved effective and powerful (Dathathri et al., 2019).

In this paper, we formulate this problem as a topic-aware dialogue response generation task, aiming to generate informative and topic-related responses that can engage the users. More specifically, we design a three-stage iteration mechanism for the topic-aware response generation task. We generate the coarse response given historical dialogue context and previous topics first, then we require the model to explicitly predict corresponding topics, and then we concatenate the generated coarse response at the first step and the predicted topics at the second step as input to generate a final refined topic-related response. Thus, the model is forced to learn a joint distribution of topics and related responses by optimizing for these three objectives simultaneously.

- We formulate a traditional response generation problem as a topic-aware generation problem and propose a joint framework that can learn topic prediction and dialogue response generation simultaneously.
- We design a topic refine mechanism to control the generation of dialogue response. Our ablation study confirms it can help to generate more informative and topic-related responses, leading to better performance.
- We evaluate our model on two different datasets which consist of two application scenarios: medical auto-diagnosis and conversational recommendation, and we achieve new state-of-the-art performance on both datasets and demonstrate that joint distribution and topic refinement is effective but the understanding ability of GPT2 still needs to be improved.

2 Problem Definition

Given a dialogue $d = \{u^1, u^2, u^3, ..., u^n\}$, a corresponding speaker role path $sr = \{s^1, s^2, s^3, ..., s^n\}$ and its corresponding topic path $tp = \{tw^1, tw^2, tw^3, ..., tw^n\}$ where $s \in R$, $tw \in T$. R and T are pre-defined speaker sets and topic sets. An utterance at ith time step can be expressed by (u^i, s^i, tw^i) which represents the sentence, the speaker, and the topics included in this sentence. tw^i consists of multiple topics or zero topic and each topic is expressed by several words. The problem then can be defined as: given a ith historical dialogue context, speaker role and topic path, $d_i^{n-1} = \{u_i^1, ..., u_i^{n-1}\}$, $sr_i^{n-1} = \{s_i^1, ..., s_i^{n-1}\}$, $tp_i^{n-1} = \{tw_i^1, ..., tw_i^{n-1}\}$, find the next topic and generate related responses.

$$y^* = \arg\max_{\theta} p(r^n, tw^n | d^{n-1}, tp^{n-1}, sr^{n-1}) \tag{1}$$

where r^n and tw^n stand for the response and corresponding topics at turn n respectively,. User profile information $p = \{p_1, p_2, ..., p_k\}$ is often provided as additional input, which consists of k sentences to express personal information such as interest. Thus, the objective changes accordingly:

$$y^* = \arg\max_{\theta} p(r^n, tw^n | d^{n-1}, tp^{n-1}, sr^{n-1}, p) \tag{2}$$

Different from other methods, we divide the whole problem into three sub-problems (see the section below). Our objective can be formulated as the following joint distribution:

$$y^* = \arg\max_{\theta} p(r_1^n|d^{n-1}, sr^{n-1}, tp^{n-1})$$
$$p(tw^n|d^{n-1}, sr^{n-1}, tp^{n-1})$$
$$p(r_2^n|d^{n-1}, sr^{n-1}, tp^{n-1}, (r_1^n, tw^n)) \quad (3)$$

where $p(r_1^n|d^{n-1}, sr^{n-1}, tp^{n-1})$ generate the relatively abbreviated response first, then $p(tw^n|d^{n-1}, sr^{n-1}, tp^{n-1})$ predict the corresponding topics at turn n, and finally, the model refines the abbreviated response r_1^n by maximizing $p(r_2^n|d^{n-1}, sr^{n-1}, tp^{n-1}, (r_1^n, tw^n))$ with the first response r_1^n and corresponding predicted topics tw^n as additional input, which leads to more informative and topic-related response r_2^t.

3 Model

Our model can be divided into three different parts: 1) Stage-One: Response Generation and 2) Topic Prediction; and 3) Stage-Two: Topic Refinement, which corresponds (a), (b), (c) shown in Figure 2 respectively. More details can be checked in the following subsections 3.1, 3.2, and 3.3.

3.1 Stage-One: Response Generation

We formulate the response generation problem using conditional language models e.g. GPT (Radford et al., 2019). Given many dialogues $D = \{d_1, d_2, d_3, ..., d_m\}$, ith dialogue d contains serval training samples $(r^n, tw^n|d^{n-1}, sr^{n-1}, tp^{n-1})$ from different turn n, our objective here is to build a statistical model parameterized by θ to maximize $p_\theta(r^n|d^{n-1}, tp^{n-1}, sr^{n-1})$. Since here we use autoregressive language models to take account of the sequential structure of the response, we need to decompose the joint probability of r^n using the chain rule as follows:

$$p_\theta(r^n|d^{n-1}, tp^{n-1}, sr^{n-1}) = \prod_{t=1}^{T} p_\theta(r_t^n|I) \quad (4)$$

where I stands for $(r_{<t}^n, d^{n-1}, tp^{n-1}, sr^{n-1})$ and $r_{<t}^n$ represents all tokens before t at turn n. The objective of stage one is performed by maximizing the loglikelihood (MLE) of the conditional probabilities in (4) over the entire training dataset:

$$L_{one} = -\sum_{m=1}^{|D|}\sum_{n=1}^{|d|}\sum_{t=1}^{T} logp_\theta(r_t^{m,n}|r_{<t}^{m,n}, \mathcal{H}_m) \quad (5)$$

where $r_{m,n}^t$ is tth token of nth resposne of mth dialogue in training dataset, $\mathcal{H}_m$ represents $(d^{m,n}, tp^{m,n}, sr^{m,n})$ before current response.

3.2 Topic Prediction

Given the historical $\mathcal{H}_m$ of m^{th} dialogue [1], we need not only to generate a suitable response but also to predict the correct topic. Some prior works solve this problem by predicting the topic first and then generating the response (Liu et al., 2020a; Zhou et al., 2020). In this section, different from these works, we propose a framework to jointly learn this task with dialogue response generation task as shown in Figure 2. There are two methods to predict the corresponding topics: (1) BERT-Based Prediction, and (2) GPT-Based Prediction.

3.2.1 BERT-Based Prediction.

Consistent with previous work in text classification (Chen et al., 2019a), we utilize the embedding h_1 of first token $[CLS]$ from BERT (Devlin et al., 2019) to predict the topics, followed by a $softmax$ layer.

$$f(x) = \text{softmax}(Wh_1 + b) \quad (6)$$

3.2.2 GPT-Based Prediction.

We adapt GPT2DoubleHeads model (Wolf et al., 2020) to perform the prediction followed (Wolf et al., 2019), since there are two heads: language modeling head and the classification head in the model while the latter one can be used to classify the input dialogue information. Besides that, the shared parameters of GPT may benefit both topic prediction and response generation tasks.

It is noted that there are two types of classification in topic prediction task: *multi-class classification* and *multi-label classification*, owing to the unique characteristic and differences of two datasets: MedDG (Liu et al., 2020a) and TG-ReDial (Liu et al., 2020b). For a *multi-class classification* problem, the global optimization can be reached by minimizing cross-entropy loss defined as follow:

$$L_{topic} = -\sum_{c=1}^{K} y_c log(p_c|\mathcal{H}_m) \quad (7)$$

For a *multi-label classification* problem, it is usually formulated as a sequence of binary decision problems which are optimized by:

[1] It is noted that we do not use $r_{<t}^n$ as input information here.

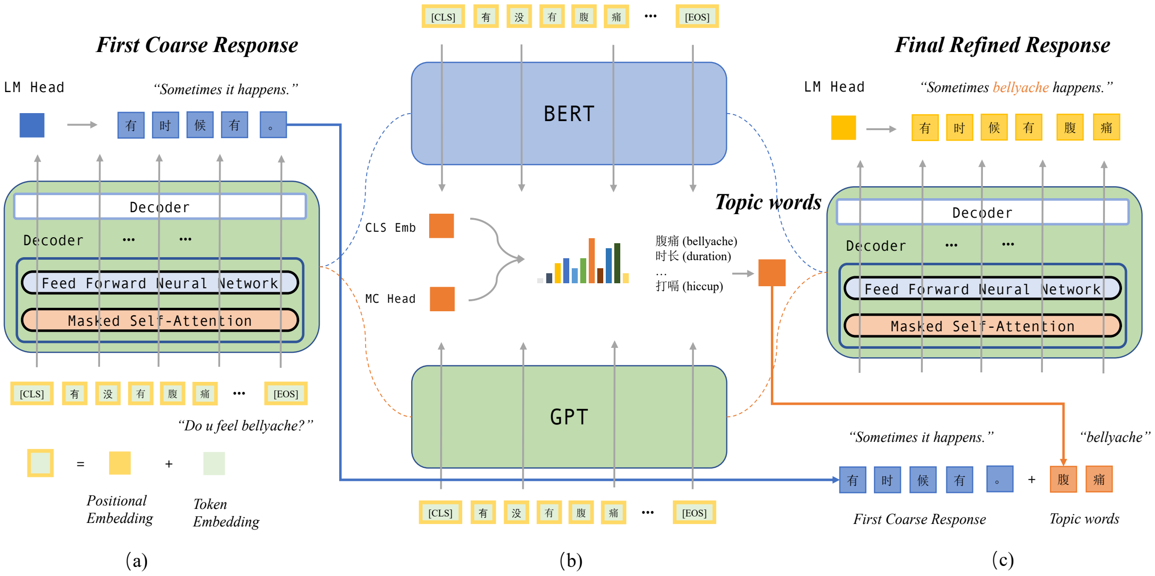


Figure 2: TopicRefine: Joint Framework of Our Proposed Model, which consists of three different modules (a) Stage-One: Response Generation (b) Topic Prediction (c) Stage-Two: Topic Refinement. The (b) module can be implemented by two methods: BERT and GPT, we utilize Stage-One (GPT) and Stage-Two (GPT) to represent the framework with GPT as the backbone for all three modules (orange dashed line), and Stage-Two (BERT) to replace GPT with BERT for (b) module (blue dashed line) in later experiment section.

$$L_{topic} = -\sum_{c=1}^{K} y_c log(p_c|\mathcal{H}_m) + (1-y_c) log(p_c|\mathcal{H}_m) \tag{8}$$

3.3 Stage-Two: Topic Refinement

To generate a more informative and topic-related response, we introduce the *topicRefine* mechanism that refines the generated response condition on the predicted topic [2], as shown in Figure 2 (c).

The refine decoder receives the first generated response r_1^n from the stage-one module and the predicted topic tw^n from the Topic Prediction module as input and outputs a refined response r_2^n. More specifically, we utilize $< topic >$ to indicate the position of topics, so the input can be represented as $\{[CLS], w_r^1, w_r^2, ..., w_r^n, < topic >, w_t^1, w_t^2, ..., w_t^n, < topic >\}$ where $r_1^n = [w_r^1, w_r^2, ..., w_r^n]$, $tw^n = [w_t^1, w_t^2, ..., w_t^n]$. The learning objective is formulated as:

$$L_{refine} = -\sum_{m=1}^{|D|}\sum_{n=1}^{|d|}\sum_{t=1}^{T} logp_\theta(r_t^{m,n}|r_{<t}^{m,n}, \mathcal{H}_m, tw^n) \tag{9}$$

where Eq 9 is similar with Eq 5 except the introduced topic information tw^n here. The parameters are shared by all three modules unless we state otherwise.

3.4 Training Objective

The learning objective of our model is the sum of three parts, jointly trained using the "teacher-forcing" algorithm. During training, we feed the ground-truth response only in stage-one and stage-two and minimize the following objective.

$$L_{model} = L_{one} + L_{topic} + L_{refine} \tag{10}$$

At test time, we choose the predicted word by $y^* = argmax_y p(y|x)$ at each time step, and we use greedy search to generate both the response and refined response.

4 Experiment

In this section, we will introduce datasets and baselines first, and then presents implementation details and evaluation metrics of our proposed framework.

4.1 Datasets

MedDG (Liu et al., 2020a) A large-scale high-quality medical dialogue dataset that contains 12 types of common diseases, more than 17k conversation, and 160 different topics consisting of symptoms and attributes. Noted the topic-prediction task here is a multi-label classification problem.

TG-ReDial (Zhou et al., 2020) consists of 10000 two-party dialogues between the user and a recommender in the movie domain which explicitly

[2] If there are k topics predicted by module b, then we simply concatenate all of them together

incorporates topic paths to enforce natural semantic transitions towards recommendation scenario. For topic-prediction task here, it is a multi-class classification problem. The details of these two datasets can be found in Table 1.

Dataset	MedDG	TG-ReDial
Task Domain	Task-oriented	Recommendation
Language	Chinese	Chinese
Classification Type	Multi-Label	Multi-Class
Dialogue Domain	Medical	Movie
♯ Dialogues	17864	10000
♯ Utterances	385951	129392
♯ Topics	160	2571

Table 1: Statistics of Two Datasets

4.2 Baselines

Seq2Seq. (Sutskever et al., 2014) is a classical attention-based sequence-to-sequence model which builds on top of vanilla RNN encoder and decoder.

HRED. (Serban et al., 2016) extends the traditional RNN encoder by stacking two RNNs in a hierarchical way: one at the word level and one at the utterance level. It is frequently used as a dialogue encoder.

GPT2. (Radford et al., 2019) is a strong baseline for response generation task which demonstrates powerful performance in many related works. It is noted all three methods mentioned above can utilize topic information as additional input which concatenates with utterance in the dialogue. We use **Seq2Seq-Topic**, **HRED-Topic** and **GPT-Topic** to represent these methods respectively.

Redial (Li et al., 2018) is proposed especially for conversational recommendation systems by utilizing an auto-encoder for the recommendation.

KBRD (Chen et al., 2019b) stands for Knowledge-Based Recommendaer Dialog System, which combines the advantages of recommendation system and dialogue generation system.

Transformer (Vaswani et al., 2017) applies a Transformer-based encoder-decoder framework to generate proper responses.

TG-RG (Zhou et al., 2020) is current state-of-the-art method comes with the release of dataset. It predicts the topic first and then generates the response.

4.3 Variants of Our Framework

GPT2DH. The method removes the refinement stage from our framework and jointly trains the response generation and topic prediction tasks (i.e. a and b module in Figure 2) based on the GPT2DoubleHeads model. In this way, the training objective changes to $L_{model} = L_{one} + L_{topic}$ without L_{refine}. We called this method GPT2DH to represent GPT2DoubleHeads (Wolf et al., 2020) which have two heads for classification and generation respectively.

Stage-One (GPT) and Stage-Two (GPT). As shown in Figure 2, this variant represents all three components are implemented by GPT2DoubleHeads model, while **Stage-One (GPT)** represents the first generated response r_1^n and **Stage-Two (GPT)** represents the refined response r_2^n in Equation (3).

Stage-Two (BERT). We replace GPT with BERT only for (b) module in Figure 2. The variant is designed for poor understanding capability of GPT model which leads to noisy predicted topic.

4.4 Implementation Details

We use the same settings for these two datasets. The learning rate is set as 1.5e-4, repetition penalty as 1.0, batch size as 4, warmup steps as 2000, except max context length as 500, max decode length as 50, epochs as 20 for TG-ReDial, max context length as 600, max decode length as 100, epochs as 10 for MedDG. We use ADAMW (Loshchilov and Hutter, 2019) to train the model. We emphasize that the role path information is missing in the test data of MedDG. Thus we only use dialogue and topic information in the experiment to keep consistent with test data. It is important to note that our methods do not pre-train on any other big corpus, we just load the parameters provided by (Wolf et al., 2020) and directly fine-tune on the target dataset.

4.5 Evaluation Metrics

For the sake of fair comparison, we adopt the same evaluation metrics as the original two papers (Liu et al., 2020a) and (Zhou et al., 2020). For MedDG, we report BLEU-1, BLEU-4, and Topic-F1 for response generation task, and Precision, Recall, and F1 score for the topic prediction task. For TG-ReDial, we calculate BLEU-1, BLEU2, and BLEU3 for generation and Hit@1, Hit@3, Hit@5 for prediction. It is noted that Topic-F1 requires the topic words appears exactly in the generated response at MedDG dataset.

5 Result and Analysis

In this section, we evaluated the proposed TopicRefine framework at two datasets MedDG and TG-ReDial respectively. And then we further investigate the effects of different response lengths and provide an analysis of human evaluation for dialogue response generation task. At the last, we also investigate the understanding capability of the GPT model in these two datasets.

5.1 Main Result

Model	BLEU-1	BLEU-4	Topic-F1	Avg
Seq2Seq	26.12	14.21	12.63	17.65
Seq2Seq-Topic	35.24	19.20	16.73	23.72
HRED	31.56	17.28	12.18	20.34
HRED-Topic	38.66	21.19	16.58	25.48
GPT2	29.35	14.47	9.17	17.66
GPT2-Topic	30.87	16.56	17.08	21.50
Stage-Two (GPT)	**45.12**	**27.49**	5.40	26.00
Stage-Two (BERT)	44.49	24.62	**17.94**	**29.02**
Stage-One (GPT)	43.86	24.62	11.36	26.61
GPT2DH	43.93	24.35	11.91	26.73

Table 2: Dialogue response generation at MedDG dataset. It is notes that “-Topic" methods use the ground truth topic information in the dataset.

Model	BLEU-1	BLEU-2	BLEU-3
Redial	0.177	0.028	0.006
KBRD	0.223	0.028	0.009
Transformer	0.283	0.068	0.033
GPT2-Topic	0.278	0.064	0.031
TG-RG	0.282	0.067	0.033
Stage-Two (GPT)	0.293	0.085	0.042
Stage-Two (BERT)	**0.294**	**0.086**	**0.043**
Stage-One (GPT)	0.284	0.082	0.041
GPT2DH	0.288	0.086	0.041

Table 3: Recommendation Response Generation at TG-ReDial dataset. It is notes that “-Topic" methods use the ground truth topic information in the dataset.

Table 2 and Table 3 demonstrates the performance of baselines and our proposed framework in both MedDG and TG-ReDial dataset respectively. Our topicRefine framework outperforms the previous state-of-the-art models at both datasets (i.e. GPT2-Topic model at MedDG and TG-RG model at TG-ReDial). More specifically, Stage-Two (GPT) reaches better BLEU score and Stage-Two (BERT) achieves higher Topic-F1 score at MedDG, owing to the existence of noisy topic in former method. Consistent with MedDG dataset, our method gets better performance no matter in Stage-Two (GPT) or Stage-Two (BERT) as shown in Table 3. BLEU-1, BLEU-2, and BLEU-3 all have been improved by different degrees. Another interesting finding is that when explicitly concatenating topic words with dialogue utterances, the GPT-Topic model achieves a higher topic-f1 score, whereas the Stage-Two (GPT) model achieves a lower topic-f1 score, indicating the effectiveness of simply concatenating topic words and the noisy prediction results by GPT.

5.2 Ablation Study

To further investigate the effectiveness of our proposed framework, we add some variants of our proposed framework (i.e. Stage-One (GPT) and GPT2DH) as ablation study. As shown in Table 2 and Table 3, Stage-One (GPT) and GPT2DH achieve comparable results. On the one hand, compared with previous state-of-the-art models, GPT2DH demonstrate more powerful capability which shows the effectiveness of joint learning by incorporating topic prediction. Besides, any Stage-Two model reaches higher BLEU scores than GPT2DH which demonstrate the effectiveness of refine mechanism (i.e. L_{refine}). On the other hand, Stage-Two (GPT) outperforms Stage-One (GPT) in BLEU score (45.12 vs 43.86) but Topic-F1 score (5.40 vs 11.36). We argue that the model tends to generate more topic-related words instead of a specific topic word in the response. This is reasonable since the model is optimized to generate a more informative and topic-related response rather than a specific word.

5.3 Effects of Response Length

To evaluate the impact of different ground-truth response length, we compare the average BLEU score between our model and previous state-of-the-art model (i.e. GPT2-Topic and TG-RG) in MedDG and TG-ReDial respectively. As shown in Figure 3 and Figure 4, our model reaches better performance when the length of golden response is greater than 20 (occupies about 47.6% and 81.9% of test set respectively). As the golden length increases, our improvements also get boosted, which is more obvious at TG-ReDial dataset.

5.4 Generated Sample

Table **??** (See Appendix due to page limit) given some generated response at both datasets. To sum-

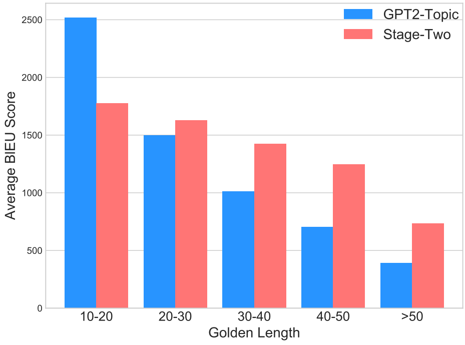


Figure 3: Average BLEU score of MedDG for different golden length

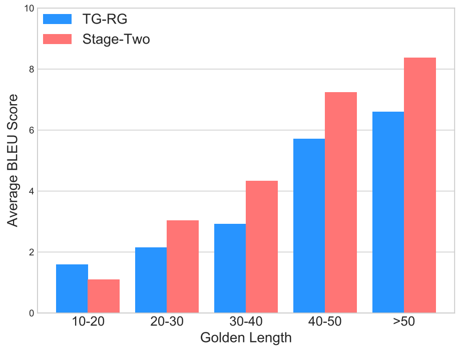


Figure 4: Average BLEU score of TG-ReDial for different golden length

marize, our generated result has the following features:

- For MedDG, since we drop the information of the speaker role path during training and the dialogue between the doctor and the patient is not alternate, some generated responses may represent the perspective of the patient.
- For TG-ReDial, there are some meaningless repeated characters in the result of Stage-One. For example, "。" and "这个电" (this movie) appears twice in response generated by Stage-One. Stage-Two can alleviate this problem by incorporating topic refinement.
- Our Stage-Two model can generate more informative responses conditioned on given topics. Taking the sample of TG-ReDial in Table **??** as an example. For the topic of "memories", the response of ground truth is just a rhetorical question, while the response of our model not only grasps this topic but also recommends one specific movie name related to this topic, which suggests that our model is able to ground multi-turn dialogue generation to some specific topics and tends to be more informative with respect to context.

Model	MedDG		TG-ReDial	
	I	F	I	F
Human	6.99	6.28	7.40	7.28
Baseline	6.18	5.51	6.20	5.69
One	6.32	4.81	6.62	5.66
Two	6.57	6.13	7.30	6.42

Table 4: The result of human evaluation. I and F represent *Information* and *Fluency* respectively. The baseline represents previous sota model *GPT2-Topic* and *TG-RG* in MedDG and TG-ReDial dataset respectively. One represents *Stage-One (GPT)* and Two represents *Stage-Two (GPT)*

5.5 Human Evaluation

To perform human evaluation, we randomly select 50 examples from the outputs of the previous sota model, and our *Stage-One (GPT)* and *State-Two (GPT)* method. The annotators are required to assign two scores for each sentence according to two criteria: (1) information and (2) fluency, ranging from 0 to 10. *information* measures which sentence contains more information (e.g. less repetition). *Fluency* measures which sentence is more proper as a response to a given dialogue context. The evaluation results are calculated by averaging these two scores of all sentences.

Table 4 demonstrates the result of human evaluation. Generally, the score at TG-ReDial dataset is relatively higher than score in MedDG dataset. We attribute this to the MedDG dataset necessitates more expert knowledge and contains many terminologies. Besides that, there is still a large gap between generated response and human response, especially at fluency criteria. In detail, the Stage-One (GPT) performs better than baseline models at information but worse at fluency. Stage-Two (GPT) model gets better scores in both information and fluency criteria than Stage-One (GPT) model and baseline.

5.6 Understanding of GPT Model

Model	P	R	F1
BERT	14.48	**32.95**	20.13
Stage-Two (GPT)	**22.22**	11.16	14.88

Table 5: Result of topic prediction task (multi-label classification) at MedDG dataset

Model	Hit@1	Hit@3	Hit@5
BERT	**0.7651**	**0.8023**	**0.8189**
Stage-Two (GPT)	0.5640	0.7931	0.8122

Table 6: Result of topic prediction task (multi-class classification) at TG-ReDial dataset

Table 5 and Table 6 demonstrate the performance of topic prediction task at MedDG and TG-ReDial datasets respectively. It is obvious that BERT (Devlin et al., 2019) demonstrates more strong understanding ability than GPT (Wolf et al., 2020) model. However, the comparable performance of Hit@3 and Hit@5 between BERT and GPT in Table 6 clearly demonstrates the latter's high understanding potential. The unlocking of potential necessitates a more meticulously designed algorithm or architecture (Dathathri et al., 2019; Liu et al., 2021).

6 Related Work

6.1 Topic-aware Dialogue System

Data-driven, knowledge-grounded dialogue system (Zhou et al., 2018b; Tuan et al., 2019; Zhao et al., 2020) attracts much attention due to the release of large pre-trained language models such as GPT2 (Radford et al., 2019) and DialoGPT (Zhang et al., 2019b). According to different types of knowledge, previous works can be clustered into the following categories: (1) attributes (Zhou et al., 2018a; Zhang et al., 2018a; Xu et al., 2019) (2) persona (Li et al., 2016; Zheng et al., 2019; Wu et al., 2020a; Zhang et al., 2018b) (3) external knowledge graph such as commonsense knowledge (Tuan et al., 2019; Yang et al., 2019; Moon et al., 2019).

Most of previous works for topic-aware dialogue system (Xing et al., 2017; Dziri et al., 2019; Yang et al., 2019; Huo et al., 2020) utilize attention mechanism on all topics at the decode stage to bias the generation probability. (Tang et al., 2019) proposes a structured approach that introduces coarse-grained keywords to control the intended content of system responses and (Xu et al., 2020) proposes Topic-Aware Dual-attention Matching (TADAM) Network to select suitable response but all of their systems are retrieval-based.

6.2 Refine Mechanism

Refine mechanism has been proved to be a effective and compelling technique in both natural language understanding and generation tasks (Zhang et al., 2019a; Wu et al., 2020b; Song et al., 2021). For natural language understanding, (Wu et al., 2020b) design a novel two-pass iteration mechanism to handle the uncoordinated slots problem caused by conditional independence of non-autoregressive model, in which the model utilizes *B-label* output from first phase as input at second phase. For natural language generation, (Zhang et al., 2019a) use refine mechanism to generate refined summary which firstly applies BERT as decoder. Recently, a novel BERT-over-BERT (BoB) model is proposed to solve response generation task and consistency understanding simultaneously (Song et al., 2021). In this paper, we utilize *topicRefine* framwork to build a topic-aware multi-turn end-to-end dialogue system, aiming to generate informative and topic-related dialogue response.

7 Conclusion and Future Work

In this paper, we propose a joint framework with a topic refinement mechanism to solve the topic-aware multi-turn end-to-end dialogue generation problem based on the auto-regressive language model – GPT2 (Wolf et al., 2020). More specifically, we design a three-pass mechanism to jointly learn topic prediction and dialogue response generation tasks, aiming to generate an informative and topic-related response to engage users. Comprehensive experiments demonstrate that our method outperforms previous state-of-the-art models on both MedDG (Liu et al., 2020a) and TG-ReDial (Liu et al., 2020b) datasets, which verifies that the effectiveness of joint learning and refinement mechanism. We will investigate more refined techniques in our future work.

Acknowledgements

This project is supported by a Research Grant, Project ID 4055160, CUHK.

References

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 7622–7634, Online. Association for Computational Linguistics.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019a. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019b. Towards knowledge-based recommender dialog system. 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)*, pages 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970, Online. Association for Computational Linguistics.

Sangwoo Cho, Kaiqiang Song, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2020. Better highlighting: Creating sub-sentence summary highlights. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6282–6300, Online. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. 2018. Topic-to-essay generation with neural networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4078–4084. ijcai.org.

Demian Gholipour Ghalandari and Georgiana Ifrim. 2020. Examining the state-of-the-art in news timeline summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1322–1334, Online. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

P. Huo, Y. Yang, J. Zhou, C. Chen, and L. He. 2020. Terg: Topic-aware emotional response generation for chatbot. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Bei Li, Ziyang Wang, Hui Liu, Yufan Jiang, Quan Du, Tong Xiao, Huizhen Wang, and Jingbo Zhu. 2020. Shallow-to-deep training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 995–1005, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9748–9758.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020a. Meddg: A large-scale medical consultation dataset for building medical dialogue system.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020b. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational ai: The science behind the alexa prize.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. 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)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020a. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 53–65, Online. Association for Computational Linguistics.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020b. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1932–1937, Online. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357. AAAI Press.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5416–5426, Florence, Italy. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2020. Topic-aware multi-turn dialogue modeling. *arXiv preprint arXiv:2009.12539*.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2002–2012, Florence, Italy. Association for Computational Linguistics.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019a. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797, Hong Kong, China. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018a. 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, Melbourne, Australia. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019. A pre-training based personalized dialogue generation model with persona-sparse data.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2020. A reinforced generation of adversarial examples for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3486–3497, Online. Association for Computational Linguistics.