Diversifying Dialogue Generation with Non-Conversational Text

Hui Su[1]*, **Xiaoyu Shen**[2]*
Sanqiang Zhao[3], **Xiao Zhou**[1], **Pengwei Hu**[4], **Randy Zhong**[1], **Cheng Niu**[1] **and Jie Zhou**[1]
[1]Pattern Recognition Center, Wechat AI, Tencent Inc, China
[2]MPI Informatics & Spoken Language Systems (LSV), Saarland Informatics Campus
[3]University of Pittsburgh [4]The Hong Kong Polytechnic University, Hong Kong
aaronsu@tencent.com,xshen@mpi-inf.mpg.de

Abstract

Neural network-based sequence-to-sequence (seq2seq) models strongly suffer from the low-diversity problem when it comes to open-domain dialogue generation. As bland and generic utterances usually dominate the frequency distribution in our daily chitchat, avoiding them to generate more interesting responses requires complex data filtering, sampling techniques or modifying the training objective. In this paper, we propose a new perspective to diversify dialogue generation by leveraging *non-conversational* text. Compared with bilateral conversations, non-conversational text are easier to obtain, more diverse and cover a much broader range of topics. We collect a large-scale non-conversational corpus from multi sources including forum comments, idioms and book snippets. We further present a training paradigm to effectively incorporate these text via iterative back translation. The resulting model is tested on two conversational datasets and is shown to produce significantly more diverse responses without sacrificing the relevance with context.

1 Introduction

Seq2seq models have achieved impressive success in a wide range of text generation tasks. In open-domain chitchat, however, people have found the model tends to strongly favor short, generic responses like "I don't know" or "OK" (Vinyals and Le, 2015; Shen et al., 2017a). The reason lies in the extreme one-to-many mapping relation between every context and its potential responses (Zhao et al., 2017; Su et al., 2018). Generic utterances, which can be in theory paired with most context, usually dominate the frequency distribution in the dialogue training corpus and thereby pushes the model to blindly produce these safe, dull responses (Su et al., 2019b; Csáky et al., 2019)

Conversational Text	
Context (Translation)	暗恋的人却不喜欢我 The one I have a crush on doesn't like me.
Response	摸摸头 Head pat.
Non-Conversational Text	
Forum Comments	暗恋这碗酒，谁喝都会醉啊 Crush is an alcoholic drink, whoever drinks it will get intoxicated.
Idiom	何必等待一个没有结果的等待 Why wait for a result without hope
Book Snippet	真诚的爱情之路永不会是平坦的 The course of true love never did run smooth (From *A Midsummer Night's Dream*)

Table 1: A daily dialogue and non-conversational text from three sources. The contents of non-conversational text can be potentially utilized to enrich the response generation.

Current solutions can be roughly categorized into two classes: (1) Modify the seq2seq itself to bias toward diverse responses (Li et al., 2016a; Shen et al., 2019a). However, the model is still trained on the *limited dialogue corpus* which restricts its power at covering broad topics in open-domain chitchat. (2) Augment the training corpus with extra information like structured world knowledge, personality or emotions (Li et al., 2016b; Dinan et al., 2019), which requires *costly human annotation*.

In this work, we argue that training only based on conversational corpus can greatly constrain the usability of an open-domain chatbot system since many topics are not easily available in the dialogue format. With this in mind, we explore a cheap way to diversify dialogue generation by utilizing large amounts of *non-conversational text*. Compared with bilateral conversations, non-conversational text covers a much broader range of topics, and can be easily obtained without further human annotation from multiple sources like forum comments, idioms and book snippets. More importantly, non-conversational text are usually *more interesting and contentful* as they are written to convey some spe-

*Equal contribution.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7087–7097
July 5 - 10, 2020.

cific personal opinions or introduce a new topic, unlike in daily conversations where people often *passively* reply to the last utterance. As can be seen in Table 1, the response from the daily conversation is a simple comfort of "Head pat". Non-conversational text, on the contrary, exhibit diverse styles ranging from casual wording to poetic statements, which we believe can be potentially utilized to enrich the response generation.

To do so, we collect a large-scale corpus containing over 1M non-conversational utterances from multiple sources. To effectively integrate these utterances, we borrow the back translation idea from unsupervised neural machine translation (Sennrich et al., 2016; Lample et al., 2018b) and treat the collected utterances as unpaired responses. We first pre-train the forward and backward transduction model on the parallel conversational corpus. The forward and backward model are then iteratively tuned to find the optimal mapping relation between conversational context and non-conversational utterances (Cotterell and Kreutzer, 2018). By this means, the content of non-conversational utterances is gradually distilled into the dialogue generation model (Kim and Rush, 2016), enlarging the space of generated responses to cover not only the original dialogue corpus, but also the wide topics reflected in the non-conversational utterances.

We test our model on two popular Chinese conversational datasets weibo (Shang et al., 2015a) and douban (Wu et al., 2017). We compare our model against retrieval-based systems, style-transfer methods and several seq2seq variants which also target the diversity of dialogue generation. Automatic and human evaluation show that our model significantly improves the responses' diversity both semantically and syntactically without sacrificing the relevance with context, and is considered as most favorable judged by human evaluators [1].

2 Related Work

The tendency to produce generic responses has been a long-standing problem in seq2seq-based open-domain dialogue generation (Vinyals and Le, 2015; Li et al., 2016a). Previous approaches to alleviate this issue can be grouped into two classes.

The first class resorts to modifying the seq2seq architecture itself. For example, Shen et al. (2018a); Zhang et al. (2018b) changes the training objective to mutual information maximization and rely on continuous approximations or policy gradient to circumvent the non-differentiable issue for text. Li et al. (2016d); Serban et al. (2017a) treat open-domain chitchat as a reinforcement learning problem and manually define some rewards to encourage long-term conversations. There is also research that utilizes latent variable sampling (Serban et al., 2017b; Shen et al., 2018b, 2019b), adversarial learning (Li et al., 2017; Su et al., 2018), replaces the beam search decoding with a more diverse sampling strategy (Li et al., 2016c; Holtzman et al., 2019) or applies reranking to filter generic responses (Li et al., 2016a; Wang et al., 2017). All of the above are still trained on the original dialogue corpus and thereby cannot generate out-of-scope topics.

The second class seeks to bring in extra information into existing corpus like structured knowledge (Zhao et al., 2018; Ghazvininejad et al., 2018; Dinan et al., 2019), personal information (Li et al., 2016b; Zhang et al., 2018a) or emotions (Shen et al., 2017b; Zhou et al., 2018). However, corpus with such annotations can be extremely costly to obtain and is usually limited to a specific domain with small data size. Some recent research started to do dialogue style transfer based on personal speeches or TV scripts (Niu and Bansal, 2018; Gao et al., 2019; Su et al., 2019a). Our motivation differs from them in that we aim at enriching general dialogue generation with abundant non-conversational text instead of being constrained on one specific type of style.

Back translation is widely used in unsupervised machine translation (Sennrich et al., 2016; Lample et al., 2018a; Artetxe et al., 2018) and has been recently extended to similar areas like style transfer (Subramanian et al., 2019), summarization (Zhao et al., 2019) and data-to-text (Chang et al., 2020). To the best of our knowledge, it has never been applied to dialogue generation yet. Our work treats the context and non-conversational text as unpaired source-target data. The back-translation idea is naturally adopted to learn the mapping between them. The contents of non-conversational text can then be effectively utilized to enrich the dialogue generation.

3 Dataset

We would like to collect non-conversational utterances that stay close with daily-life topics and can

[1] Code and dataset available at https://github.com/chin-gyou/Div-Non-Conv

Resources	Size	Avg. length
Comments	781,847	21.0
Idioms	51,948	18.7
Book Snippets	206,340	26.9

Table 2: Statistics of Non-Conversational Text.

be potentially used to augment the response space. The utterance should be neither too long nor too short, similar with our daily chitchats. Therefore, we collect data from the following three sources:

1. Forum comments. We collect comments from zhihu [2], a popular Chinese forums. Selected comments are restricted to have more than 10 likes and less than 30 words [3].

2. Idioms. We crawl idioms, famous quotes, proverbs and locutions from several websites. These phrases are normally highly-refined and graceful, which we believe might provide a useful augmentation for responses.

3. Book Snippets. We select top 1,000 favorite novels or prose from wechat read [4]. Snippets highlighted by readers, which are usually quintessential passages, and with the word length range 10-30 are kept.

We further filter out sentences with offensive or discriminative languages by phrase matching against a large blocklist. The resulting corpus contains over 1M utterances. The statistics from each source are listed in Table 2.

4 Approach

Let $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$ denote the parallel conversational corpus. X_i is the context and Y_i is the corresponding response. $\mathcal{D}_T = \{T_1, T_2, \ldots, T_M\}$ denotes our collected corpus where T_i is a non-conversational utterance. As the standard seq2seq model trained only on $\mathcal{D}$ tends to generate over-generic responses, our purpose is to diversify the generated responses by leveraging the non-conversational corpus $\mathcal{D}_T$, which are semantically and syntactically much richer than responses contained in $\mathcal{D}$. In the following section, we first go through several baseline systems, then introduce our proposed method based on back translation.

4.1 Retrieval-based System

The first approach we consider is a retrieval-based system that considers all sentences contained in $\mathcal{D}_T$ as candidate responses. As the proportion of generic utterances in $\mathcal{D}_T$ is much lower than that in $\mathcal{D}$, the diversity will be largely improved. Standard retrieval algorithms based on context-matching (Wu et al., 2017; Bartl and Spanakis, 2017) fail to apply here since non-conversational text does not come with its corresponding context. Therefore, we train a backward seq2seq model on the parallel conversational corpus $\mathcal{D}$ to maximize $p(X_i|Y_i)$. The score assigned by the backward model, which can be seen as an estimation of the point-wise mutual information, is used to rank the responses (Li et al., 2016a) [5].

The major limitation of the retrieval-based system is that it can only produce responses from a finite set of candidates. The model can work well only if an appropriate response already exists in the candidate bank. Nonetheless, due to the large size of the non-conversational corpus, this approach is a very strong baseline.

4.2 Weighted Average

The second approach is to take a weighted average score of a seq2seq model trained on $\mathcal{D}$ and a language model trained on $\mathbf{D}_T$ when decoding responses. The idea has been widely utilized on domain adaptation for text generation tasks (Koehn and Schroeder, 2007; Wang et al., 2017; Niu and Bansal, 2018). In our scenario, basically we hope the generated responses could share the diverse topics and styles of the non-conversational text, yet stay relevant with the dialogue context. The seq2seq model $S2S$ is trained on $\mathcal{D}$ as an indicator of how relevant each response is with the context. A language model $\mathcal{L}$ is trained on $\mathcal{D}_T$ to measure how the response matches the domain of $\mathcal{D}_T$. The decoding probability for generating word w at time step t is assigned by:

$$p_t(w) = \alpha S2S_t(w) + (1 - \alpha)L_t(w) \quad (1)$$

[2] https://www.zhihu.com

[3] The posts are usually very long, describing a specific social phenomenon or news event, so building parallel conversational corpus from post-comment pairs is difficult. Nonetheless, these high-liked comments are normally high-quality themselves and can be used to augment the response space.

[4] https://weread.qq.com/

[5] The backward seq2seq model measures the context relevance better than forward models since the latter highly biases generic utterances (Li et al., 2016a; Zhang et al., 2018b)

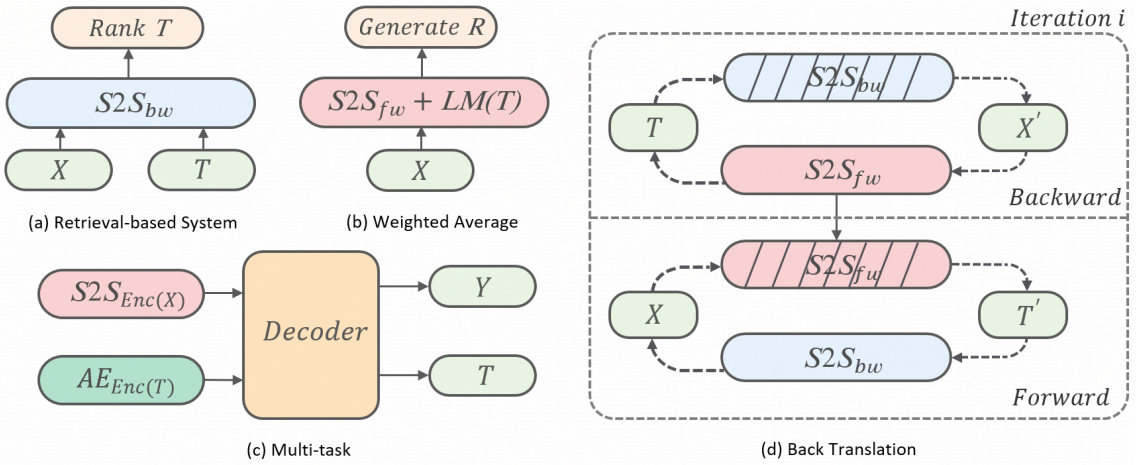

Figure 1: Comparison of four approaches leveraging the non-conversational text. $S2S_{fw}$, $S2S_{bw}$ and LM indicate the forward, backward seq2seq and language model respectively. (d) visualizes the process of one iteration for the back translation approach. Striped component are not updated in each iteration.

where α is a hyperparameter to adjust the balance between the two. Setting $\alpha = 1$ will make it degenerate into the standard seq2seq model while $\alpha = 0$ will totally ignore the dialoge context.

4.3 Multi-task

The third approach is based on multi-task learning. A seq2seq model is trained on the parallel conversational corpus $\mathcal{D}$ while an autoencoder model is trained on the non-parallel monologue data $\mathcal{D}_T$. Both models share the decoder parameters to facilitate each other. The idea was first experimented on machine translation in order to leverage large amounts of target-side monolingual text (Luong et al., 2016; Sennrich et al., 2016). Luan et al. (2017) extended it to conversational models for speaker-role adaptation. The intuition is that by tying the decoder parameters, the seq2seq and autoencoder model can learn a shared latent space between the dialogue corpus and non-conversational text. When decoding, the model can generate responses with features from both sides.

4.4 Back Translation

Finally, we consider the back translation technique commonly used for unsupervised machine translation (Artetxe et al., 2018; Lample et al., 2018a). The basic idea is to first *initialize* the model properly to provide a good starting point, then iteratively perform *backward* and *forward* translation to learn the correspondence between context and unpaired non-conversational utterances.

Initialization Unlike unsupervised machine translation, the source and target side in our case come from the same language, and we already have a parallel conversational corpus $\mathcal{D}$, so we can get rid of the careful embedding alignment and autoencoding steps as in Lample et al. (2018b). For the initialization, we simply train a forward and backward seq2seq model on $\mathcal{D}$. The loss function is:

$$\mathbb{E}_{X_i,Y_i\sim\mathcal{D}} - \log P_f(Y_i|X_i) - \log P_b(X_i|Y_i) \quad (2)$$

where P_f and P_b are the decoding likelihood defined by the forward and backward seq2seq model respectively. We optimize Eq. 2 until convergence. Afterwards, the forward and backward seq2seq can learn the backbone mapping relation between a context and its response in a conversational structure.

Backward After the initialization, we use the backward seq2seq to create pseudo parallel training examples from the non-conversational text $\mathcal{D}_T$. The forward seq2seq is then trained on the pseudo pairs. The objective is to minimize:

$$\begin{aligned} &\mathbb{E}_{T_i\sim\mathcal{D}_T} - \log P_f(T_i|b(T_i)) \\ &b(T_i) = \arg\max_u P_b(u|T_i) \end{aligned} \quad (3)$$

where we approximate the $\arg\max$ function by using a beam search decoder to decode from the backward model $P_b(u|T_i)$. Because of the non-differentiability of the $\arg\max$ operator, the gradient is only passed through P_f but not P_b [6].

As P_b is already well initialized by training on the parallel corpus $\mathcal{D}$, the back-translated pseudo

[6] As also noted in Lample et al. (2018b), backpropagating further through P_b brings no improvement.

(**Initialization**) Train by minimizing Eq. 2 until convergence;
for *i=1 to N* **do**
  (**Backward**) Train by minimizing Eq. 3 until convergence;
  (**Forward**) Train by minimizing Eq. 4 until convergence;
end

Algorithm 1: Model Training Process

pair $\{b(T_i), T_i\}$ can roughly follow the typical human conversational patterns. Training P_f on top of them will encourage the forward decoder to generate utterances in the domain of T_i while maintaining coherent as a conversation.

Forward The forward translation follows a similar step as back translation. The forward seq2seq P_f translates context into a response, which in return form a pseudo pair to train the backward model P_b. The objective is to minimize:

$$\begin{aligned} &\mathbb{E}_{X_i \sim \mathcal{D}} - \log P_b(X_i | f(X_i)) \\ &f(X_i) = \arg\max_v P_f(v | X_i) \end{aligned} \tag{4}$$

where the $\arg\max$ function is again approximated with a beam search decoder and the gradient is only backpropagated through P_b. Though X_i has its corresponding Y_i in $\mathcal{D}$, we drop Y_i and instead train on forward translated pseudo pairs $\{X_i, f(X_i)\}$. As P_f is trained by leveraging data from $\mathcal{D}_T$, $f(X_i)$ can have superior diversity compared with Y_i.

The encoder parameters are shared between the forward and backward models while decoders are separate. The backward and forward translation are iteratively performed to close the gap between P_f and P_b (Hoang et al., 2018; Cotterell and Kreutzer, 2018). The effects of non-conversational text are strengthened after each iteration. Eventually, the forward model will be able to produce diverse responses covering the wide topics in $\mathcal{D}_T$. Algorithm 1 depicts the training process.

5 Experiments

5.1 Datasets

We conduct our experiments on two Chinese dialogue corpus Weibo (Shang et al., 2015b) and Douban (Wu et al., 2017). Weibo [7] is a popular Twitter-like microblogging service in China, on which a user can post short messages, and other users make comment on a published post. The post-comment pairs are crawled as short-text conversations. Each utterance has 15.4 words on average and the data is split into train/valid/test subsets with 4M/40k/10k utterance pairs. Douban [8] is a Chinese social network service where people can chat about different topics online. The original data contains 1.1M multi-turn conversations. We split them into two-turn context-response pairs, resulting in 10M train, 500k valid and 100K test samples.

5.2 General Setup

For all models, we use a two-layer LSTM (Hochreiter and Schmidhuber, 1997) encoder/decoder structure with hidden size 500 and word embedding size 300. Models are trained with Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.15. We set the batch size as 256 and use the gradients clipping of 5. We build out vocabulary with character-based segmentation for Chinese. For non-Chinese tokens, we simply split by space and keep all unique tokens that appear at least 5 times. Utterances are cut down to at most 50 tokens and fed to every batch. We implement our models based on the OpenNMT toolkit (Klein et al., 2017) and other hyperparameters are set as the default values.

5.3 Compared Models

We compare our model with the standard seq2seq and four popular variants which were proposed to improve the diversity of generated utterances. All of them are trained only on the parallel conversational corpus:

Standard The standard seq2seq with beam search decoding (size 5).

MMI The maximum mutual information decoding which reranks the decoded responses with a backward seq2seq model (Li et al., 2016a). The hyperparameter λ is set to 0.5 as suggested. 200 candidates per context are sampled for re-ranking

Diverse Sampling The diverse beam search strategy proposed in Vijayakumar et al. (2018) which explicitly controls for the exploration and exploitation of the search space. We set the number of groups as 5, $\lambda = 0.3$ and use the Hamming diversity as the penalty function as in the paper.

[7] http://www.weibo.com/

[8] https://www.douban.com/group

Nucleus Sampling Proposed in Holtzman et al. (2019), it allows for diverse sequence generations. Instead of decoding with a fixed beam size, it samples text from the dynamic nucleus. We use the default configuration and set $p = 0.9$.

CVAE The conditional variational autoencoder (Serban et al., 2017b; Zhao et al., 2017) which injects diversity by imposing stochastical latent variables. We use a latent variable with dimension 100 and utilize the KL-annealing strategy with step 350k and a word drop-out rate of 0.3 to alleviate the posterior collapse problem (Bowman et al., 2016).

Furthermore, we compare the 4 approaches mentioned in §4 which incorporate the collected non-conversational text:

Retrieval-based (§4.1) Due to the large size of the non-conversational corpus, exact ranking is extremely slow. Therefore, we first retrieve top 200 matched text with elastic search based on the similarity of Bert embeddings (Devlin et al., 2019). Specifically, we pass sentences through Bert and derive a fixed-sized vector by averaging the outputs from the second-to-last layer (May et al., 2019) [9]. The 200 candidates are then ranked with the backward score [10].

Weighted Average (§4.2) We set $\lambda = 0.5$ in eq. 1, which considers context relevance and diversity with equal weights.

Multi-task ((§4.3)) We concatenate each context-response pair with a non-conversational utterance and train with a mixed objective of seq2seq and autoencoding (by sharing the decoder).

Back Translation (§4.4) We perform the iterative backward and forward translation 4 times for both datasets. We observe the forward cross entropy loss converges after 4 iterations.

6 Results

As for the experiment results, we report the automatic and human evaluation in §6.1 and §6.2 respectively. Detailed analysis are shown in §6.3 to elaborate the differences among model performances and some case studies.

[9] https://github.com/hanxiao/bert-as-service

[10] This makes it similar to MMI reranking, whose 200 candidates are from seq2seq decodings instead of top-matched non-conversational utterances.

6.1 Automatic Evaluation

Evaluating dialogue generation is extremely difficult. Metrics which measure the word-level overlap like BLEU (Papineni et al., 2002) have been widely used for dialogue evaluation. However, these metrics do not fit into our setting well as we would like to diversify the response generation with an external corpus, the generations will inevitably differ greatly from the ground-truth references in the original conversational corpus. Though we report the BLEU score anyway and list all the results in Table 3, it is worth mentioning that the BLEU score itself is by no means a reliable metric to measure the quality of dialogue generations.

Diversity Diversity is a major concern for dialogue generation. Same as in (Li et al., 2016a), we measure the diversity by the ratio of distinct unigrams (**Dist-1**) and bigrams (**Dist-2**) in all generated responses. As the ratio itself ignores the frequency distribution of n-grams, we further calculate the entropy value for the empirical distribution of n-grams (Zhang et al., 2018b). A larger entropy indicates more diverse distributions. We report the entropy of four-grams (**Ent-4**) in Table 3. Among models trained only on the conversational corpus, the standard seq2seq performed worst as expected. All different variants improved the diversity more or less. Nucleus sampling and CVAE generated most diverse responses, especially Nucleus who wins on 6 out of the 8 metrics. By incorporating the non-conversational corpus, the diversity of generated responses improves dramatically. The retrieval-based system and our model perform best, in most cases even better than human references. This can happen as we enrich the response generation with external resources. The diversity would be more than the original conversational corpus. Weighted-average and multi-task models are relatively worse, though still greatly outperforming models trained only on the conversational corpus. We can also observe that our model improves over standard seq2seq only a bit after one iteration. As more iterations are added, the diversity improves gradually.

Relevance Measuring the context-response relevance automatically is tricky in our case. The typical way of using scores from forward or backward models as in Li and Jurafsky (2017) is not suitable as our model borrowed information from extra resources. The generated responses are out-of-scope

Metrics	Weibo					Douban				
Model	BLEU-2	Dist-1	Dist-2	Ent-4	Adver	BLEU-2	Dist-1	Dist-2	Ent-4	Adver
STANDARD	0.0165	0.018	0.050	5.04	0.30	0.0285	0.071	0.206	7.55	0.19
MMI	0.0161	0.025	0.069	5.98	0.42	0.0263	0.143	0.363	7.60	**0.31**
DIVERSE	0.0175	0.019	0.054	6.20	0.38	0.0298	0.130	0.358	7.51	0.25
NUCLEUS	**0.0183**	**0.027**	**0.074**	**7.41**	**0.43**	**0.0312**	0.141	0.402	**7.93**	0.30
CVAE	0.0171	0.023	0.061	6.63	0.36	0.0287	**0.169**	**0.496**	7.80	0.29
RETRIEVAL	0.0142	**0.198**	**0.492**	**12.5**	0.13	**0.0276**	**0.203**	**0.510**	**13.3**	**0.17**
WEIGHTED	**0.0152**	0.091	0.316	9.26	0.22	0.0188	0.172	0.407	8.73	0.14
MULTI	0.0142	0.128	0.348	8.98	**0.27**	0.0110	0.190	0.389	8.26	0.16
BT (ITER=1)	**0.0180**	0.046	0.171	7.64	0.19	**0.0274**	0.106	0.313	8.16	0.15
BT (ITER=4)	0.0176	**0.175**	**0.487**	**11.2**	**0.35**	0.0269	**0.207**	**0.502**	**11.0**	**0.25**
HUMAN	-	0.171	0.452	9.23	0.88	-	0.209	0.514	11.3	0.85

Table 3: Automatic evaluation on Weibo and Douban datasets. Upper areas are models trained only on the conversational corpus. Middle areas are baseline models incorporating the non-conversational corpus. Bottom areas are our model with different number of iterations. Best results in every area are **bolded**.

for the seq2seq model trained on only on the conversational corpus and thus would be assigned very low scores. Apart from the BLEU-2 score, we further evaluate the relevance by leveraging an adversarial discriminator (Li et al., 2017). As has been shown in previous research, discriminative models are generally less biased to high-frequent utterances and more robust against their generative counterparts (Lu et al., 2017; Luo et al., 2018). The discriminator is trained on the parallel conversational corpus distinguish correct responses from randomly sampled ones. We encode the context and response separately with two different LSTM neural networks and output a binary signal indicating relevant or not [11]. The relevance score is defined as the success rate that the model fools the adversarial classifier into believing its generations (**Adver** in Table 3). The retrieval-based model, who generates the most diverse generations, achieve the lowest score as for relevance with context. The restriction that it can only select from a set of fixed utterances do affect the relevance a lot [12]. Note that *the discriminator is also trained on the same bilateral conversational corpus, putting our model into a naturally disadvantageous place due to the incorporation of out-of-scope non-conversational text.* Nonetheless, our model still achieves competitive relevance score even compared with models trained only on the conversational corpus. This suggests our model does learn the proper patterns in human conversations instead of randomly synthesizing diverse generations.

Metrics	Weibo			Douban		
Model	Rel	Inter	Flu	Rel	Inter	Flu
STANDARD	0.32	0.11	0.76	0.26	0.13	0.82
NUCLEUS	**0.46**	0.19	**0.78**	0.38	0.21	**0.83**
RETRIEVAL	0.12	0.35	-	0.09	0.32	-
WEIGHTED	0.19	0.14	0.52	0.15	0.17	0.46
MULTI	0.25	0.21	0.70	0.22	0.23	0.66
BT (ITER=4)	0.43	**0.37**	0.77	**0.39**	**0.48**	0.80

Table 4: Human Evaluation Results

6.2 Human Evaluation

Apart from automatic evaluations, we also employed crowdsourced judges to evaluate the quality of generations for 500 contexts of each dataset. We focus on evaluating the generated responses regarding the (1) relevance: if they coincide with the context (**Rel**), (2) interestingness: if they are interesting for people to continue the conversation (**Inter**) and (3) fluency: whether they are fluent by grammar (**Flu**) [13]. Each sample gets one point if judged as yes and zero otherwise. Each pair is judged by three participants and the score supported by most people is adopted. The averaged scores are summarized in Table 4. We compare the standard seq2seq model, nucleus sampling which performs best among all seq2seq variants, and the four approaches leveraging the non-conversational text. All models perform decently well as for fluency except the weighted average one. The scores for diversity and relevance generally correlate well with the automatic evaluations. Overall the back-translation model are competitive with respect to fluency and relevance, while generating much more interesting responses to human evaluators. It also significantly outperforms the other three baseline

[11] In our experiment, the discriminator performs reasonably well in the 4 scenarios outlined in Li et al. (2017) and thus can be considered as a fair evaluation metric.

[12] The fact that we only rank on 200 most similar utterances might also affect. We tried increasing the size to 1,000 but observe no tangible improvement. The candidate size required for a decent relevance score can be unbearably large.

[13] We do not evaluate the retrieval-based model for the fluency score as the retrieved utterances are fluent by construct.

Context	一直单身怎么办 (Being always single, what should I do?)
Response	勇敢一点多去加好友啊 (Be brave and add more people to friends.)
Generation	[Iteration 0]: 不知道该怎么办 (I don't know what to do.)
	[Iteration 1]: 单身不可怕，单身不可怕 (Being single is nothing, being single is nothing.)
	[Iteration 4]: 斯人若彩虹，遇上方知有 (Every once in a while you find someone who's iridescent, and when you do, nothing will ever compare.)

Table 5: Example of response generation in different iterations.

approaches in its capability to properly make use of the non-conversational corpus.

6.3 Analysis

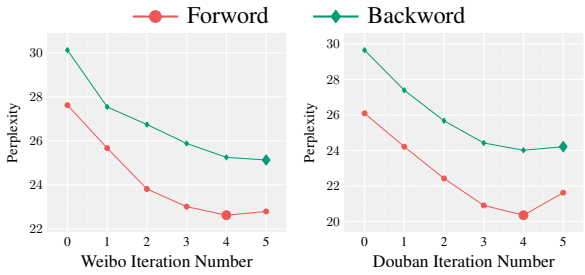

Figure 2: Change of validation loss across iterations.

Effect of Iterative Training To show the importance of the iterative training paradigm, we visualize the change of the validation loss in Figure 2 [14]. The forward validation loss is computed as the perplexity of the forward seq2seq on the pseudo context-response pairs obtained from the backward model, vice versa for backward loss. It approximately quantifies the KL divergence between them two (Kim and Rush, 2016; Cotterell and Kreutzer, 2018). As the iteration goes, the knowledge from the backward model is gradually distilled into the forward model. The divergence between them reaches the lowest point at iteration 4, where we stop our model. Table 5 further displays examples for different iterations. Iteration 0 generates mostly generic responses. Iteration 1 starts to become more diverse but still struggle with fluency and relevance. In the final iteration, it can learn to incorporate novel topics from the non-conversational text yet maintaining the relevance with context.

[14]Iteration 0 means before the iteration starts but after the initialization stage, equal to a standard seq2seq.

CXT	最近又长胖了 Fleshing out again recently.
NS	我也是这样的 Me too.
BT	哈哈哈莫非已经**胖若两人**了 hahaha already **as fat as two people**?
CXT	爱一个人真的不能跟她表白吗? Why loving someone but cannot confess?
NS	不一定的 Not necessarily.
BT	爱一个人**不难，难的是**放下一个人。 **To** love **is easy, to** give up **is hard.**

Table 6: Context (CXT), example generations from neucleus sampling (NS) and back-translation (BT). Novel words and syntax patterns are **highlighted**.

Diversity of Generation We find the back translation model can generate *both semantically and syntactically* novel responses. Some examples are shown in Table 6. To find semantically novel responses, we segment them into phrases and find those containing novel phrases that do not exist on the conversational corpus. As in the first example of Table 6, the word 胖若两人 only exists in the non-conversational corpus. The model successfully learnt its semantic meaning and adopt it to generate novel responses. It is also common that the model learns frequent syntax structures from the non-conversational corpus. In the second example, it learnt the pattern of "To ... is easy, to ... is hard", which appeared frequently in the non-conversational corpus, and utilized it to produce novel responses with the same structure. Note that both generations from the BT model *never appear exactly in the non-conversational corpus*. It must generate them by correctly understanding the meaning of the phrase components instead of

memorizing the utterances verbally.

7 Conclusion and Future Work

We propose a novel way of diversifying dialogue generation by leveraging non-conversational text. To do so, we collect a large-scale corpus from forum comments, idioms and book snippets. By training the model through iterative back translation, it is able to significantly improve the diversity of generated responses both semantically and syntactically. We compare it with several strong baselines and find it achieved the best overall performance. The model can be potentially improved by filtering the corpus according to different domains, or augmenting with a retrieve-and-rewrite mechanism, which we leave for future work.

Acknowledgments

We thank anonymous reviewers for valuable comments. Xiaoyu Shen is supported by IMPRS-CS fellowship. The work is partially funded by DFG collaborative research center SFB 1102.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. *ICLR*.

Alexander Bartl and Gerasimos Spanakis. 2017. A retrieval-based dialogue system utilizing utterance and context embeddings. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1120–1125. IEEE.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Ernie Chang, David Ifeoluwa Adelani, Xiaoyu Shen, and Vera Demberg. 2020. Unsupervised pidgin text generation by pivoting english data and self-training. *arXiv preprint arXiv:2003.08272*.

Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv preprint arXiv:1806.04402*.

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *ICLR*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1814–1823.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, pages 224–227.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. *ICLR*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Jiwei Li and Dan Jurafsky. 2017. Neural net models for open-domain discourse coherence. *EMNLP*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016c. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016d. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Jiwei Li, Will Monroe, Tianlin Shi, Sėbastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 605–614.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *ICLR*.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017a. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015a. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015b. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Xiaoyu Shen, Youssef Oualil, Clayton Greenberg, Mittul Singh, and Dietrich Klakow. 2017a. Estimation of gap between current language models and human performance. *Proc. Interspeech 2017*, pages 553–557.

Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. 2018a. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316–4327.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017b. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 504–509.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018b. Improving variational encoder-decoders in dialogue generation. *AAAI*, pages 5456–5463.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019a. Select and attend: Towards controllable content selection in text generation. *arXiv preprint arXiv:1909.04453*.

Xiaoyu Shen, Yang Zhao, Hui Su, and Dietrich Klakow. 2019b. Improving latent alignment in text summarization by generalizing the pointer generator. 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 3753–3764.

Feng-Guang Su, Aliyah R Hsu, Yi-Lin Tuan, and Hung-Yi Lee. 2019a. Personalized dialogue response generation learned from monologues. *Proc. Interspeech 2019*, pages 4160–4164.

Hui Su, Xiaoyu Shen, Pengwei Hu, Wenjie Li, and Yun Chen. 2018. Dialogue generation with gan. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019b. Improving multi-turn dialogue modelling with utterance rewriter. *arXiv preprint arXiv:1906.07004*.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text style transfer. *ICLR*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David J Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *AAAI*, pages 7371–7379.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. 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.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 654–664.

Yang Zhao, Xiaoyu Shen, Wei Bi, and Akiko Aizawa. 2019. Unsupervised rewriter for multi-sentence compression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2240.

Yang Zhao, Xiaoyu Shen, Hajime Senuma, and Akiko Aizawa. 2018. A comprehensive study: Sentence compression with linguistic knowledge-enhanced gated neural network. *Data & Knowledge Engineering*, 117:307–318.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.