Non-Autoregressive Neural Dialogue Generation

Qinghong Han[1*], Yuxian Meng[1*], Fei Wu[2] and Jiwei Li[1]
[1] ShannonAI
[2] Department of Computer Science and Technology, Zhejiang University
{qinghong_han, jiwei_li}@shannonai.com, wufei@zju.edu.cn

Abstract

Maximum Mutual information (MMI), which models the bidirectional dependency between responses (y) and contexts (x), i.e., the forward probability $\log p(y|x)$ and the backward probability $\log p(x|y)$, has been widely used as the objective in the SEQ2SEQ model to address the dull-response issue in open-domain dialog generation. Unfortunately, under the framework of the SEQ2SEQ model, direct decoding from $\log p(y|x) + \log p(x|y)$ is infeasible since the second part (i.e., $p(x|y)$) requires the completion of target generation before it can be computed, and the search space for y is enormous. Empirically, an N-best list is first generated given $p(y|x)$, and $p(x|y)$ is then used to rerank the N-best list, which inevitably results in non-globally-optimal solutions.

In this paper, we propose to use non-autoregressive (non-AR) generation model to address this non-global optimality issue. Since target tokens are generated independently in non-AR generation, $p(x|y)$ for each target word can be computed as soon as it's generated, and does not have to wait for the completion of the whole sequence. This naturally resolves the non-global optimal issue in decoding. Experimental results demonstrate that the proposed non-AR strategy produces more diverse, coherent, and appropriate responses, yielding substantive gains in BLEU scores and in human evaluations.[1]

1 Introduction

Open-domain neural dialogue generation (Vinyals and Le, 2015; Sordoni et al., 2015; Li et al., 2016a; Mou et al., 2016; Serban et al., 2016a; Asghar et al., 2016; Mei et al., 2016; Serban et al., 2016e,b,d; Baheti et al., 2018; Wang et al., 2018; Ghazvininejad et al., 2018; Zhang et al., 2018; Gao et al., 2019) treats dialog contexts (x) as sources,and responses (y) as targets and uses the encoder-decoder model (Sutskever et al., 2014; Vaswani et al., 2017b) as the backbone to generate responses. SEQ2SEQ models offer the promise of scalability and language-independence, along with the capacity to capture contextual dependencies semantic and syntactic relations between sources and targets.

One of key issues with the SEQ2SEQ structure is that it exhibits a strong tendency to generate dull, trivial or non-committal responses (e.g., *I don't know* or *I'm OK*) regardless of the input, which has been observed by many recent works (Li et al., 2016a; Sordoni et al., 2015; Serban et al., 2016c; Niu and Bansal, 2020). Various strategies (Li et al., 2016a; Vijayakumar et al., 2016; Baheti et al., 2018; Niu and Bansal, 2020) have been proposed to address this issue, , one of the most widely used of which is to replace the MLE objective in the SEQ2SEQ training with the maximum mutual information objective (MMI for short) (Li et al., 2016a). MMI models the bidirectional dependency between responses (y) and contexts (x). It takes the form of the linear combination of the forward probability $\log p(y|x)$ and the backward probability $\log p(x|y)$. The intuition behind MMI is straightforward: it is easy to predict a dull response given any context, but hard to predict the context given a dull response since the context that corresponds to a dull response could be anything.

Unfortunately, under the framework of the SEQ2SEQ model, direct decoding from $\log p(y|x) + \log p(x|y)$ is infeasible since the second part (i.e., $p(x|y)$) requires the completion of target generation before $p(x|y)$ can be computed, and the search space for y is huge. Empirically, an N-best list is first generated given $p(y|x)$, and $p(x|y)$ is then used to rerank the

[1] Qinghong and Yuxian contribute equally to this work.

N-best list. Due to the fact that beam search lacks for diversity in the beam: candidates often differ only by punctuation or minor morphological variations, with most of the words overlapping, this reranking strategy inevitably results in non-globally-optimal solutions. Some strategies have been proposed to alleviate this non-global-optimality issue, such as generating a more diverse N-best list (Li et al., 2016c; Gu et al., 2017; Vijayakumar et al., 2016), or using reinforcement learning to estimate the future score of $p(x|y)$ (Li et al., 2017a), which help alleviate the non-globally-optimal issue, but cannot fully address it.

Non-autoregressive (non-AR) generation (Gu et al., 2018; Ma et al., 2019; Lee et al., 2018) provides resolution to the non-global-optimality issue. Under the formalization of non-AR generation, target tokens y_t are generated independently, which enables $p(x|y_t)$ to be computed as soon as y_t is generated. This naturally resolves the non-global optimal issue in decoding. We conduct experiments on the widely used Opensubtitle dataset and experimental results demonstrate that the proposed strategy produces more diverse, coherent, and appropriate responses, yielding substantive gains in BLEU scores and in human evaluations.

The rest of this paper is organized as follows: Section 2 and section 3 present related work and background knowledge respectively. The propose model is described in Section 4. Experimental results and ablation studies are detailed in Section 5 and 6, followed by a brief conclusion in Section 7.

2 Related Work

2.1 Neural Dialogue Generation

End-to-end neural approaches for dialogue generation use SEQ2SEQ architectures (Sutskever et al., 2014; Vaswani et al., 2017b) as the backbone to generate syntactically fluent and meaningful responses, providing the flexibility to capture contextual semantics between source contexts and target responses. Recent studies have endowed these models with the ability to model contexts (Sordoni et al., 2015; Serban et al., 2016e,b; Tian et al., 2017; Lewis et al., 2017), generating coherent and personalized responses (Li et al., 2016b; Zhao et al., 2017; Shao et al., 2017; Xing et al., 2017; Zhang et al., 2018; Bosselut et al., 2018), generating uttterances with different attributes or topics (Wang et al., 2017; Niu and Bansal, 2018) and interacting fluently with humans (Ghazvininejad et al., 2018; Zhang et al., 2019; Adiwardana et al., 2020).

2.2 Diverse Decoding

One major issue with SEQ2SEQ systems is their propensity to select dull, non-committal responses regardless of the input, for which many diverse decoding algorithms have been proposed to tackle this problem (Li et al., 2016a; Li and Jurafsky, 2016; Vijayakumar et al., 2016; Cho, 2016; Kulikov et al., 2018; Kriz et al., 2019; Ippolito et al., 2019). Li et al. (2016a) proposed to use Maximum Mutual Information (MMI) as the objective function in neural dialog models. MMI models use both the forward probability $p(y|x)$ and the backward probability $p(x|y)$ to better capture the contextual relations between the source and target sequences. Li and Jurafsky (2016) introduced a Beam Search diversification heuristic to discourage sequences from sharing common roots, implicitly resulting in diverse sequences. Vijayakumar et al. (2016) improved upon Li and Jurafsky (2016) and presented Diverse Beam Search, which formalizes beam search as an optimization problem and augments the objective with a diversity term. Cho (2016) introduced Noisy Parallel Approximate Decoding, a method encouraging diversity by adding small amounts of noise to the hidden state of the decoder at each step, instead of manipulating the probabilities outputted from the model. Kulikov et al. (2018) attempted to explore larger beam search space by running beam search many times, where the states explored by subsequent beam searches are restricted based on the intermediate states explored by previous iterations. These works have pushed dialogue models to generate more interesting and diverse responses that are both high-quality and meaningful.

2.3 Non-Autoregressive Sequence Generation

Besides diverse responses, another problem for these dialogue generation models is their autoregressive generation strategy that decodes words one-by-one, making it extremely slow to execute on long sentences, especially on conditions where multi-turn dialogue often appears (Adiwardana et al., 2020). One solution is to use non-autoregressive sequence generation methods, which has recently aroused general interest in the community of neural machine translation (NMT) (Gu et al., 2018; Lee et al., 2018; Ma et al., 2019; Sun et al., 2019; Shu et al., 2019; Bao et al., 2019). Gu et al. (2018) proposed to alleviate latency by using fertility during inference in autoregressive Seq2Seq NMT systems, which led to a ∼15 times speedup to traditional autoregressive methods, whereas the performance degrades rapidly. Lee et al. (2018); Ma et al. (2019); Shu et al. (2019) proposed to use latent variables to model intermediate word alignments between source and target sequence pairs and mitigate the trade-off between decoding speed and performance. Bao et al. (2019) pointed out position information is crucial for non-autoregressive models and thus proposed to explicitly model position as latent variables. Sun et al. (2019) incorporated CRF into non-autoregressive models to enhance local dependencies during decoding. This work is greatly inspired by these advances in non-autoregressive sequence generation.

3 Background

3.1 Autoregressive Seq2Seq Models

An encoder-decoder model (Sutskever et al., 2014; Vaswani et al., 2017b; Bahdanau et al., 2014) defines the probability of a target sequence $Y = \{y_1, y_2, ..., y_{L_y}\}$, which is a response in the context of dialogue generation, given a source sequence $X = \{x_1, x_2, ..., x_{L_x}\}$, where where L_x and L_y are the length of the source and target sentence respectively.

An autoregressive encoder-decoder model decomposes the distribution over a target sequence $\boldsymbol{y} = \{y_1, \cdots, y_{L_y}\}$ into a chain of conditional probabilities:

$$p_{\text{AR}}(\boldsymbol{y}|\boldsymbol{x};\phi) = \prod_{t=1}^{L_y+1} \log p(y_t|y_{0:t-1}, x_{1:L_x}; \theta)$$
$$= \prod_{t=1}^{m} \frac{\exp(f(h_{t-1}, e_{y_t}))}{\sum_{y'} \exp(f(h_{t-1}, e_{y'}))} \quad (1)$$

with y_0 being the special $< BOS >$ token and y_{L_y+1} being the special $< EOS >$ token. The probability of generating a token y_t depends on all tokens in the source X, and all its previous tokens $y_{0:t-1}$ in Y. The concatenation of X and $y_{0:t-1}$ is mapped to a representation h_{t-1} using LSTMs (Sutskever et al., 2014), CNNs (Gehring et al., 2017) or transformers (Vaswani et al., 2017b). e_{y_t} denotes the representation for y_t.

During decoding, the algorithm terminates when the $< EOS >$ token is predicted. At each time step, either a greedy approach or beam search can be adopted for word prediction. Greedy search selects the token with the largest conditional probability, the embedding of which is then combined with preceding output to predict the token at the next step.

3.2 Non-Autoregressive Seq2Seq Models

3.2.1 Overview

The autoregressive generation model has two major drawbacks: it prohibits generating multiple tokens simultaneously, which leads to inefficiency in GPU usage; and erroneously generated tokens leads to error accumulation and the performance of beam search deteriorates when exposed to a larger search space (Koehn and Knowles, 2017). Non-autoregressive methods address these two issues by removing the sequential dependencies within the target sentence and generating *all* target tokens simultaneously, with the probability giving as follows:

$$p_{\text{Non-AR}}(\boldsymbol{y}|\boldsymbol{x};\phi) = \prod_{t=1}^{L_y} p(y_t|\boldsymbol{x};\phi) \quad (2)$$

Now that each target token y_t only depends on the source sentence $\boldsymbol{x}$, the full target sentence can be decoded in parallel, where `argmax` is applied to each token. A vital challenge that non-autoregressive face is the *inconsistency problem* Gu et al. (2018), which indicates the decoded sequence contains duplicated or missing tokens. Improving decoding consistency on the target side is thus crucial to Non-AR models.

4 Model

4.1 Overview

The maximum mutual information (MMI) model, proposed in (Li et al., 2016a), tries to find the response that has the largest value of mutual information with respect to the context. The form of MMI is given as follows:[2]

$$\hat{y} = \arg\max_{y} \left\{ (1-\lambda) \log p(y|x) + \lambda \log p(x|y) \right\} \quad (3)$$

[2] We refer readers to (Li et al., 2016a) for how Eq.3 is obtained.

This weighted MMI objective function can be viewed as representing a tradeoff between sources given targets (i.e., $p(x|y)$) and targets given sources (i.e., $p(y|x)$). Direct decoding from $\log(1-\lambda)p(y|x) + \lambda \log p(x|y)$ is infeasible since the second part (i.e., $p(x|y)$) requires the completion of target generation before $p(x|y)$ can be computed. Empirically, an N-best list is first generated given $p(y|x)$, and $p(x|y)$ is then used to rerank the N-best list, which inevitably results in non-globally-optimal solutions.

Here to propose to use Non-AR generation models to handle to non-globally-optimality issue. The generation of each target word y_t is independent under the non-AR formalization, and the forward probability $p(y|x)$ is given as follows:

$$\text{forward\_prob} = \prod_{t=1}^{t=L_y} p(y_t|x) \tag{4}$$

For the backward probability $p(x|y)$, which denotes the probability of generating a source sequence given a target sequence, we propose to replace it with the geometric mean of the probability of generating the source sequence given each target token, denoted as follows:

$$\text{backward\_prob} = [\prod_{t=1}^{t=L_y} p(x|y_t)]^{1/L_y} \tag{5}$$

We also use the non-AR framework to model the backward probability. Based on the independence assumption of non-AR, in which the generations of x_t are independent, Eq. 5 can be further factorized as follows:

$$\text{backward\_prob} = [\prod_{t=1}^{t=L_y} \prod_{t'=1}^{t'=L_x} p(x_{t'}|y_t)]^{1/L_y} \tag{6}$$

A close look at Equ.6 shows that it actually mimics the IBM model (Brown et al., 1993): $p(x_{t'}|y_t)$ handles the pairwise word alignment between sources and targets. Since position representations are incorporated at both the encoding and decoding stage, Eq.6 actually mimics IBM model2, where relative positions between source and target words are modeled.

Combining the forward probability in Eq. 4.2 and the backward probability in Eq.6, the full form of mutual information of Eq.3 can be rewritten as follows:

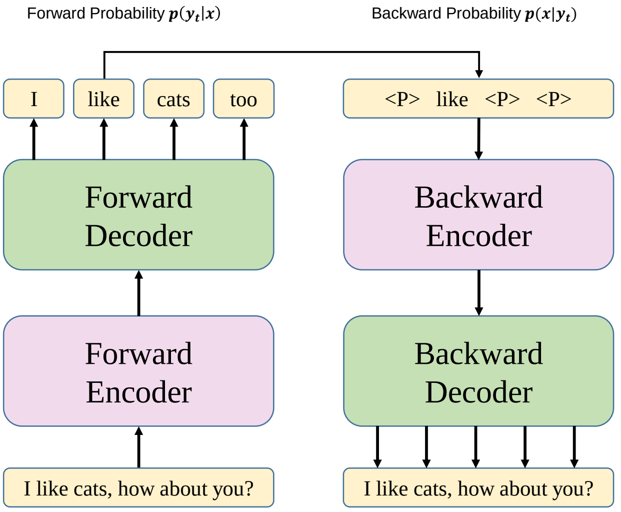
Forward Probability $p(y_t|x)$
Backward Probability $p(x|y_t)$
I
like
cats
too
<P> like <P> <P>
Forward Decoder
Backward Encoder
Forward Encoder
Backward Decoder
I like cats, how about you?
I like cats, how about you?

Figure 1: Overview of the non-auto MMI generation model.

$$\begin{aligned} L =& (1-\lambda)\sum_{t=1}^{t=L_y} \log p(y_t|x) + \frac{\lambda}{L_y}\sum_{t=1}^{t=L_y}\sum_{t'=1}^{t'=L_x} \log p(x_{t'}|y_t) \\ =& \sum_{t=1}^{t=L_y} [(1-\lambda)\log p(y_t|x) + \frac{\lambda}{L_y}\sum_{t'=1}^{t'=L_x} \log p(x_{t'}|y_t)] \end{aligned} \tag{7}$$

as can be seen, we are able to factorize the full form of the MMI objective with respect to y_t under the framework of non-AR generation. This means that the mutual information between source x and different target words y_t are independent and can be computed in parallel. Also, for each token y_t, its mutual information with respect to the source x can be readily computed as soon as y_t is generated, and we do not have to wait until the completion of the entire sequence. This naturally resolves the non-globally-optimality issue in the AR generation model. Figure 1 gives an illustration for the proposed model.

4.2 Forward Probability $p(y|x)$

We use the non-autoregressive SEQ2SEQ model as the backbone to compute $\prod_t p(y_t|x)$, which consists of two major components: the encoder and the decoder.

4.2.1 Encoder

We use transformers (Vaswani et al., 2017a) as a backbone and use a stack of $N = 6$ identical transformer blocks as the encoder. Given the source sequence $\boldsymbol{x} = \{x_1, \cdots, x_n\}$, the encoder produces its contextual representations $\boldsymbol{H} = \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\}$ from the last layer of the encoder.

4.2.2 Decoder

Target Length We first need to obtain the length of the target sequence for decoding. We follow previous works (Gu et al., 2018; Ma et al., 2019; Bao et al., 2019) to predict the length difference Δm between source and target sequences using a classifier with a range of [-20, 20]. This is accomplished by max-pooling the source embeddings into a single vector, running this through a linear layer followed by a softmax operation, as follows:

$$p(\Delta m|\boldsymbol{x}) = \text{softmax}(W_p(\text{maxpool}(\boldsymbol{H})) + b_p) \tag{8}$$

Decoder Structure The decoder also consists of $N = 6$ identical transformer blocks. The i-th position of the input $\boldsymbol{d}_i$ to the decoder is the round$(n * (i/m))$-th input's contextual representation $\boldsymbol{h}_{\text{round}(n*(i/m))}$ copied from the encoder, which is equivalent to scanning the source inputs from left to right and leads to a deterministic decoding process given the predicted target length. Both absolute and relative positional embeddings are incorporated. For relative position information, we follow Shaw et al. (2018) which produces a different learned embedding according to the offset between the "key" and "query" in the self-attention mechanism with a clipping distance k (we set $k = 4$) for relative positions. For absolute positional embeddings, we follow Radford et al. (2019) and used a learnable positional embedding $\boldsymbol{p}_t$ for position t.

Attention over Vocabulary Layer-wise attention over vocabulary is incorporated into each decoding layer to make the model aware of which token is to be generated regarding each position. More concretely, we use $\boldsymbol{Z}^{(i)}(1 \leq i \leq 6)$ to denote the contextual representations for the i-th decoder layer , and $\boldsymbol{Z}^{(0)} = \{\boldsymbol{d}_1, \cdots, \boldsymbol{d}_m\}$ to denote the input to the decoder. The intermediate token attention representation $\boldsymbol{a}_j^{(i)}$ of position $j(1 \leq j \leq m)$ in the i-th decoder layer is thus given by:

$$\boldsymbol{a}_j^{(i)} = \text{softmax}(\boldsymbol{z}_j^{(i)} \cdot \boldsymbol{W}^{\mathrm{T}}) \cdot \boldsymbol{W} \tag{9}$$

where $\boldsymbol{W}$ is the representation matrix of the token vocabulary. By doing so, each position is able to know which token is about to be decoded at the current position. The input to the next layer $\boldsymbol{Z}^{(i+1)}$ is the concatenation of the contextual representations and the intermediate token representations $[\boldsymbol{Z}^{(i)}; \boldsymbol{A}^{(i)}]$.

softmax For each position t, $p(y_t|x)$ is computed by outputting the representation for that position to a softmax function.

4.3 Backward Probability $p(x|y)$

We use the non-AR model to obtain $p(x|y_t)$.

4.3.1 Encoder

The encoder for $p(x|y_t)$ is again a stack of $N = 6$ identical transformer blocks. The input to the encoder is a text sequence with length being L_y, which is identical to the length of the target. The t-th position of the input sequence is the word y_t, with the rest being the place-holding dummy token. For each posiition, the embedding for the absolute position and the embedding for the relative position are appended.

4.3.2 Decoder

The decoder for the backward probability is the same as that of the forward probability, with the only difference being changing target y to source x.

4.4 Decoding from Mutual Information

The most commonly used decoding strategy for non-AR generation is the noisy parallel decoding strategy (NPD for short) proposed in Gu et al. (2018): a number of sequence candidates are first generated by the non-AR generation, then an AR SEQ2SEQ model is used to select the candidate that has the largest value of probability output from the AR model. Since this NPD strategy is used for the MLE objective which only concerns about the forward probability, we need to tailor it to the MMI objective. Specifically, we first generate N-best sequences based on the score of non-AR MMI function, computed from Eq.7. The final selected response is the sequence with highest AR MMI score, which is computed based on two AR SEQ2SEQ models, one to model the forward probability and the other to model the backward probability.

5 Experiments

5.1 Datasets

We use the OpenSubtitles dataset for evaluation. It's a widely used open-domain dataset, which contains roughly 60M-70M scripted lines spoken by movie characters. It has been used in a broad range of recent work on data-driven conversation This dataset does not specify which character speaks each subtitle line, which prevents us from inferring speaker turns. Following (Vinyals and Le, 2015; Li et al., 2016a), we make an assumption that each line of subtitle constitutes a full speaker turn. Although this assumption is often violated, prior work has successfully trained and evaluated neural conversation models using this corpus. In our experiments we used a preprocessed version of this dataset distributed by Li et al. (2016a).[3]

The noisy nature of the OpenSubtitle dataset renders it unreliable for evaluation purposes. We thus follow Li et al. (2016a) to use data from the Internet Movie Script Database (IMSDB)[4] for evaluation. The IMSDB dataset explicitly identifies which character speaks each line of the script. We followed protocols in (Li et al., 2016a) and randomly selected two subsets as development and test datasets, each containing 2,000 pairs, with source and target length restricted to the range of [6,18].

5.2 Baselines

Our baselines include the AR generation models (using or not using MMI) based on transformers (Vaswani et al., 2017b), with the number of encoder and decoder blocks set to 6. For the standard AR model, the value of beam size is set to 10 for decoding, and the sequence with the largest value of $p(y|x)$ is selected. For AR+MMI, we followed Li et al. (2016a), and first use $p(y|x)$ to generate an N-best list with beam-size 10. Then $p(x|y)$ is used to rerank the N-best list. λ is treated as the hyper-parameter to be tuned on the dev set.

We also implement two variant of the AR+MMI model: (1) AR+MMI+diverse (Li et al., 2016c), which uses a diverse decoding model to generate the N-best list and uses the backward probability to rerank the diverse N-best list. The diverse decoding model adds an additional term to penalize siblings in beam searchexpansions of the same parent node in the search thus favoring choosing hypotheses from diverse parents; and (2) AR+MMI+RL (Li et al., 2017a), which incorporates the critic that estimates further backward probability into decoding.

5.3 Training Details

All experiments were run using 64 Nvidia V100 GPUs with mini-batches of approximately 100K tokens. We use the same hyper-parameters for all experiments, i.e., word representations of size 1024, feed-forward layers with inner dimension 4096. Dropout rate is set to 0.2 and the number of attention heads is set to 16. Models are optimized with Adam (Kingma and Ba, 2014) using $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e8$. Differentiable scheduled sampling Goyal et al. (2017) is used to mitigate the exposure bias issue. We train models with 16-bit floating point operations. The backward model and the forward model are jointly trained with word embeddings shared.

5.4 Automatic Evaluation

For automatic evaluation, we report the results of the following metrics:

- the BLEU score following previous work. It should be noted that BLEU is not generally accepted (Liu et al., 2016) to match human evaluation in generation tasks since there are distinct ways to reply to an input.

- *distinct-1* and *distinct-2* (Li et al., 2016a): calculating the number of distinct unigrams and bigrams in generated responses scaled by total number of generated unigrams and bigrams.

- Avg.length: the average length of the generated response.

- Stopword%: the percentage of stop-words[5] of the responses generated by each model.

- Adversarial Success: the adversarial evaluation strategy proposed by Kannan and Vinyals (2017); Li et al. (2017b). Adversarial evaluation trains a discriminator (or evaluator) function to labels dialogues as machine-generated (negative) or human-generated (positive). Positive

[3] http://nlp.stanford.edu/data/OpenSubData.tar

[4] http://www.imsdb.com/

[5] Thecombinationofstopwordsinhttps://www.ranks.nl/stopwordsandpunctuations.

Model	BLEU	distinct-1	distinct-2	Avg.length	Stopword	adv succ
Human	-	16.8%	58.1%	14.2	69.8%	
AR	1.64	3.7%	9.5%	6.4	82.3%	2.7%
AR+MMI	2.10	10.6%	20.5%	7.2	76.4%	6.3%
AR+MMI+diverse	2.16	16.0%	27.3%	7.5	72.1%	6.4%
AR+MMI+RL	2.34	13.7%	25.2%	7.3	73.0%	8.0%
NonAR	1.54	8.9%	14.6%	7.1	77.9%	2.4%
NonAR+MMI	2.68	15.9%	27.0%	7.4	71.9%	9.2%

Table 1: Automatic Metrics Evaluation for Different Models.

examples are taken from training dialogues, while negative examples are decoded using generative models from a model. Adversarial success is the percentage of the generated responses that can fool the evaluator to believe that it is human-generated. We refer readers to Li et al. (2017b) for more details about the adversarial evaluation.

Results are shown in Table 1. When comparing AR with AR+MMI, AR+MMI significantly outperforms AR across all metrics, which is in line with previous findings (Li et al., 2016a). For the variants of AR+MMI, AR+MMI+diverse generates a more diverse N-best list for reranking, and thus outperforms AR+MMI; AR+MMI+RL uses lookahead strategy to estimate future backward probability, and thus outperforms AR+MMI as well. It's hard to tell which model performs better, AR or non-AR: AR performs better than non-AR for BLEU and adversarial success, but worse for the other metrics. This means comparing with AR model, non-AR model tends to generate more diverse responses, but might be less coherent. Because of the ability to handle the non-local-optimality issue, Non-AR+MMI consistently outperforms AR+MMI by a large margin across all evaluation metrics. When comparing non-AR with AR+MMI+diverse, non-AR has relatively lower *distinct* score, but significantly higher scores BLEU and adversarial success. This is because the diverse decoding strategy in AR sacrifices language model probability for diversity, and thus harms the BLEU score but promotes the diversity score. NonAR+MMI outperforms AR+MMI+RL across all metrics.

5.5 Examples

5.6 Qualitative Evaluation

We employed crowdsourced judges to provide evaluations for a random sample of 1000 items from the test set. Following protocols in Baheti et al. (2018), we assigned each output to a human judge, who were asked to score every model response on a 5-point scale (Strongly Agree, Agree, Unsure, Disagree, Strongly Disagree) on 2 categories: 1) Coherence - is the response coherent to the given source? and 2) Content Richness - does the response add new information to the conversation? Ratings were later collapsed to 3 categories (Agree, Unsure, Disagree).

The results for plausibility and content richness of different models are presented in Table 3. For dialogue coherence, the trend is that NonAR+MMI is better than AR+MMI, followed by AR and Non-AR. AR is slightly better than Non-AR. For Content Richness, the proposed NonAR+MMI is significantly better than AR+MMI, and the gap is greater than dialogue coherence. This is because the N-best list generated by the AR model tends to be dull and generic, and the reranking model in AR+MMI can help alleviate but cannot fully address this issue. The output from the AR+MMI model is thus by far less diverse than nonAR+MMI, which obtains the MMI score for each generated token.

To verify the statistical significance of the reported results, we performed a pairwise bootstrap test (Johnson, 2001; Berg-Kirkpatrick et al., 2012) to compare the difference between percentage of responses that were labeled as yes. We computed p-values for non-AR+MMI vs AR+MMI and non-AR vs AR. Regarding non-AR vs AR, we did not find a significant difference (p-value = 0.18) for coherence, but a significant difference for content richness (p-value < 0.01). For non-AR+MMI vs AR+MMI and

Input	AR+MMI	NonAR+MMI
it feels like i must have been asleep for weeks.	i m sorry that i can 't help you.	Yeah, it has been long.
where does she work?	i don 't know where she works.	she has work to do at the library.
who is in charge?	I don 't know sir.	i 'm the captain in charge!
I am off all week next week.	what time you going?	then who will be in charge ?
why can 't you just believe us?	I mean , it s not like we re gonna be together forever .	because you are not a good person.
can 't you see how they re exploiting you?	i m not a human being .	it 's not that simple.
I mean, wed have to talk to him.	I mean, he 's a good guy.	let 's do it.
i 'm sorry to detain you for so long.	i 'm sorry i can 't.	that 's fine since i 'm not leaving until tomorrow.
do you have any idea what caused the explosion ?	i have no idea.	the explosion was caused by the failure of the motor.

Table 2: Response generation: Sample responses using the diversity-promoting beam search and vanilla beam search.

Model	disagr (%)	un(%)	agr(%)
Coherence			
Human	17.4	20.8	61.8
AR	28.6	29.5	41.9
AR+MMI	25.3	27.9	46.8
AR+MMI+diverse	24.8	27.8	47.4
AR+MMI+RL	24.1	26.5	49.4
nonAR	29.9	28.7	41.4
nonAR+MMI	23.1	24.0	52.9
Content Richness			
Human	14.0	16.6	69.4
AR	38.2	30.4	31.4
AR+MMI	30.6	26.2	43.2
AR+MMI+diverse	23.9	21.3	54.8
AR+MMI+RL	26.4	24.9	48.7
NonAR	31.4	25.0	44.6
NonAR+MMI	24.2	20.5	55.3

Table 3: Human judgments for Coherence and Content Richeness of the different models.

AR+MMI+RL, we find a significant difference for both coherence (p-value < 0.01) and content richness (p-value < 0.01). For non-AR+MMI vs AR+MMI+RL, the difference for coherence is significant (p-value < 0.01), but content richness is insignificant (p-value=0.25).

5.7 Sample Responses

Sample responses are presented in Table 2. As can be seen, the nonAR+MMI tends to generate more diverse and content-rich responses. It is also interesting to see that responses from the AR+MMI model mostly start with the word "*I* ". This is because of the fact that the N-best list from the AR model lacks for diversity. The prefixes of the responses are mostly the same and the reranking process can only affect suffixes. On the contrary, for nonAR+MMI, MMI reranking is performed once a token is generated, and does not wait for the completion of the whole target sequence, leading to more diverse and appropriate responses.

5.8 Results on Machine Translation

Mutual information has been found to improve machine translation, both in the context of NMT models (Li and Jurafsky, 2016) and phrase-based MT models (Och and Ney, 2002; Shen et al., 2010). It would be interesting to see whether the proposed model can also help non-AR NMT as well. We evaluate the proposed method on the three widely used machine translation benchmark tasks (three datasets): WMT2014 De→En (4.5M sentence pairs), WMT2014 En→De, WMT2016 Ro→En (610K sentence pairs) and IWSLT2014 De→En (150K sentence pairs). We use the Transformer (Vaswani et al., 2017a) as a backbone. **Knowledge Distillation** is applied for all models. Since building SOTA non-AR MT models is out of the scope of this paper, we used the commonly used NonAR structure described in Section 4.2 as the backbone. Results are shown in Table 4. As can be seen, the incorporation of MMI model significantly improves MT performances. This shows that the proposed model has potentials to benefit a wide range of generation tasks.

6 Conclusion

In this paper, we propose to use non-autoregressive (non-AR) generation to address the non-global optimality issue for MMI in neural dialog generation. Target tokens are generated independently in non-AR generation. $p(x|y)$ for each target word can thus be computed as soon as it s generated, and does not have to wait

	WMT14 En→De	WMT14 De→En	WMT16 Ro→En
NAT (Gu et al., 2018)	17.69	20.62	29.79
iNAT (Lee et al., 2018)	21.54	25.43	29.32
FlowSeq-large (raw data) (Ma et al., 2019)	20.85	25.40	29.86
NAT (our implementation)	22.32	24.83	29.93
NAT +MMI	23.80	26.05	30.50
	(+1.48)	(+1.22)	(+0.57)

Table 4: The performances of NonAR+MMI methods on WMT14 En↔De and WMT16 Ro→En. Results from Gu et al. (2018); Lee et al. (2018); Ma et al. (2019) are copied from original papers for reference purposes.

for the completion of the whole sequence. This naturally resolves the non-global optimal issue in decoding. Experimental results demonstrate that the proposed strategy produces more diverse, coherent, and appropriate responses, yielding substantive gains in BLEU scores and in human evaluations.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Nabiha Asghar, Pasca Poupart, Jiang Xin, and Hang Li. 2016. Online sequence-to-sequence reinforcement learning for open-domain conversational agents. *arXiv preprint arXiv:1612.03929*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. *arXiv preprint arXiv:1809.01215*.

Yu Bao, Hao Zhou, Jiangtao Feng, Mingxuan Wang, Shujian Huang, Jiajun Chen, and Lei Li. 2019. Non-autoregressive transformer by position learning. *arXiv preprint arXiv:1911.10677*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Kyunghyun Cho. 2016. Noisy parallel approximate decoding for conditional recurrent language model. *arXiv preprint arXiv:1605.03835*.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

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*.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017. Differentiable scheduled sampling for credit assignment. *arXiv preprint arXiv:1704.06970*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Jiatao Gu, Kyunghyun Cho, and Victor OK Li. 2017. Trainable greedy decoding for neural machine translation. *arXiv preprint arXiv:1702.02429*.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Roger W Johnson. 2001. An introduction to the bootstrap. *Teaching Statistics*, 23(2):49–54.

Anjuli Kannan and Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198*.

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

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. 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 3137–3147, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilia Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of search and evaluation strategies in neural dialogue modeling.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proc. of NAACL-HLT*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. 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.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation.

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

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017a. Learning to decode for future success. *arXiv preprint arXiv:1701.06549*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017b. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. Flowseq: Non-autoregressive conditional sequence generation with generative flow.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2016. Coherent dialogue with attention-based language models. *arXiv preprint arXiv:1611.06997*.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6(0).

Tong Niu and Mohit Bansal. 2020. Avgout: A simple output-probability measure to eliminate dull responses. *arXiv preprint arXiv:2001.05467*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 295–302. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Iulian V Serban, II Ororbia, G Alexander, Joelle Pineau, and Aaron Courville. 2016a. Multi-modal variational encoder-decoders. *arXiv preprint arXiv:1612.00377*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16)*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016c. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016d. Generative deep neural networks for dialogue: A short review.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016e. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219, Copenhagen, Denmark. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency statistical machine translation. *Computational Linguistics*, 36(4):649–671.

Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. 2019. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Meg Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT*.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast structured decoding for sequence models. In *Advances in Neural Information Processing Systems*, pages 3011–3020.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of ICML Deep Learning Workshop*.

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, Copenhagen, Denmark. Association for Computational Linguistics.

William Yang Wang, Jiwei Li, and Xiaodong He. 2018. Deep reinforcement learning for nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–21.

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*, AAAI17, page 33513357. AAAI Press.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

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

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)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.