Another Diversity-Promoting Objective Function for Neural Dialogue Generation

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Techonology, Japan

RIKEN, Center for Advanced Intelligence Project AIP, Japan

{nakamura.ryo.nm8, sudoh, koichiro, s-nakamura}@is.naist.jp

Abstract

Although generation-based dialogue systems have been widely researched, the response generations by most existing systems have very low diversities. The most likely reason for this problem is Maximum Likelihood Estimation (MLE) with Softmax Cross-Entropy (SCE) loss. MLE trains models to generate the most frequent responses from enormous generation candidates, although in actual dialogues there are various responses based on the context. In this paper, we propose a new objective function called Inverse Token Frequency (ITF) loss, which individually scales smaller loss for frequent token classes and larger loss for rare token classes. This function encourages the model to generate rare tokens rather than frequent tokens. It does not complicate the model and its training is stable because we only replace the objective function. On the OpenSubtitles dialogue dataset, our loss model establishes a state-of-the-art DIST-1 of 7.56, which is the unigram diversity score, while maintaining a good BLEU-1 score. On a Japanese Twitter replies dataset, our loss model achieves a DIST-1 score comparable to the ground truth.

1 Introduction

Researchers have widely investigated generation-based dialogue systems and are making rapid progress in this area. However, a common problem remains: dialogue systems tend to produce such generic responses as "I don't know." Some studies have artificially promoted diversity. The diversity-promoting objective function based on Maximum Mutual Information (MMI) first addressed this kind of problem (Li et al. 2016a), and various generative model-based methods (e.g., GAN and VAE) have been proposed (Li et al. 2017; Xu et al. 2018; Olabiyi et al. 2018; Cao and Clark 2017; Zhao, Zhao, and Eskenazi 2017).

The most likely reason for this problem is Maximum Likelihood Estimation (MLE) with Softmax Cross-Entropy (SCE) loss. Although many different responses exist in actual dialogues, when you are talking to a human, MLE trains the model to generate frequent phrases in the training set, such as "I'm sorry," "I'm not sure," and "I don't know."

To solve this low diversity problem, we propose a new objective function called Inverse Token Frequency (ITF) loss, which scales loss based on the ITF at each time step. This

new function encourages the model to generate rare tokens rather than frequent tokens. ITF loss creates the following advantages:

- The ITF loss model yields state-of-the-art diversity and maintains the quality. It is also very clear, easy to understand, and sufficiently novel.

- ITF loss can be easily incorporated, and the MMI, GAN, VAE, and RL implementations become complicated because models are modified or added.

- Training with ITF loss is as stable as training with MLE, and training with GAN and RL is usually unstable and often requires pre-training with MLE.

2 Related Works

Low diversity problems in neural dialogue generation were first addressed by Li et al. (2016a) who augmented the objective function with Maximum Mutual Information (MMI). Their work promoted diversity by penalizing generic responses with an anti-language model. For sustainable dialogue generation, a reinforcement learning-based method was proposed by Li et al. (2016b). The negative cosine similarity between an input and a response was given as a reward, but the improvement of the diversity was small. Controlling output tokens by attention or an extension to LSTM cells leads to the diversity of response generation (Wen et al. 2015b; Zhou et al. 2017; Shao et al. 2017). Encoding dialog histories and external resources also promoted diversity (Ghazvininejad et al. 2017). Even though other works addressed over-generation and reranking (Wen et al. 2015a; Li et al. 2016a; Serban et al. 2017a), a model must be built that can generate a variety of sentences.

Recently, several generative model-based methods have been proposed. The Generative Adversarial Network (GAN) was proposed in image generation (Goodfellow et al. 2014) and applied to text generation (Yu et al. 2017) and dialogue generation (Li et al. 2017; Xu et al. 2018; Olabiyi et al. 2018). Currently, training with GAN for dialogue generation is very unstable and requires pre-training. Variational AutoEncoder (VAE) was also proposed in image generation (Kingma and Welling 2013) and applied to text generation (Bowman et al. 2016) and dialogue generation (Cao and Clark 2017; Zhao, Zhao, and Eskenazi 2017).

Table 1: Examples of token frequencies and corresponding weights (w) with $\lambda = 0.4$ on English OpenSubtitles dialogue. We tokenized sentences by a subword model with a 32k vocabulary using Sentencepiece (Kudo 2018), and an underscore (_) stands for a word boundary given by Sentencepiece.

index	token	freq	w	index	token	freq	w	index	token	freq	w
10	_I	1096434	0.00384	1000	_strong	2872	0.0414	10000	_cupboard	186	0.124
20	_it	383979	0.00584	2000	But	1281	0.0571	15000	_cruelty	107	0.154
50	_don	128837	0.00904	3000	rd	795	0.0692	20000	TOO	69	0.184
100	_them	54395	0.0128	4000	_print	559	0.0796	25000	_planetarium	46	0.216
500	_happen	7040	0.0289	5000	_bottles	425	0.0888	30000	ebulon	29	0.260

3 Methods

The task of response generation can be formulated as a sequence-to-sequence problem that generates a response based on given inputs. In neural dialogue generation, training with Maximum Likelihood Estimation (MLE) approximates model distribution that generates response sentence $Y = \{y_1, y_2, \ldots, y_m\}$ to a true distribution that gives target sentence $T = \{t_1, t_2, \ldots, t_m\}$. Generally, the loss function individually calculates the loss between generated token y_i and target token t_i across all token symbols. The following section describes the loss at any time step i.

3.1 Softmax Cross-Entropy Loss

Softmax Cross-Entropy (SCE) loss, which is commonly used when training a sequence-to-sequence model with MLE, is typically defined as:

$$L_{\text{sce}} = -\log\left(\frac{\exp(x_c)}{\sum_k^{|V|} \exp(x_k)}\right), \quad (1)$$

where x_k is the k-th element of $x \in \mathbb{R}^{|V|}$, which is the output of the projection layer before the softmax layer, and c is the index of the target token class. SCE loss treats each token class equally. Therefore, the generation probabilities of the frequent tokens become too large, and those of the rare tokens become too small. This problem causes the model to select only frequent tokens from an enormous number of token candidates.

3.2 Inverse Token Frequency Loss

We propose Inverse Token Frequency (ITF) loss to deal with the bias of SCE loss and to promote diversity. ITF loss is a frequency-weighted version of SCE loss:

$$L_{\text{ITF}} = w_c L_{\text{sce}} \quad (2)$$
$$w_c = \frac{1}{\text{freq}(\text{token}_c)^\lambda}, \quad (3)$$

where w_c is an element corresponding to class c in weight $w \in \mathbb{R}^{|V|}$, token_c is a token corresponding to class c, and function $\text{freq}(\text{token}_c)$ is the frequency with which token_c appears in the training set. Hyperparameter λ controls the frequency's impact. When $\lambda = 0$, the ITF loss is equivalent to the SCE loss. The distribution drawn from the softmax layer is the same for training and evaluation. Special tokens,

such as Start and End (i.e., starting and ending of sentences), are handled identically as the others. Therefore, these tokens are very small weight in the ITF loss because they appear in all of the sentences in the training set. We found no serious problems, including generating inappropriately long responses by weighting. In Table 1, we show some examples of token frequencies.

Finally, we show a code example of ITF loss implementation with PyTorch:

```
def get_weights(_lambda):
    weights = torch.zeros(vocab_size)
    for token, index in token2index.items():
        weight = 1 / (token2freq[token]**_lambda)
        weights[index] = weight
    return weights

weights = get_weights(_lambda=0.4)
itf_loss = nn.NLLLoss(weight=weights)
log_softmax = nn.LogSoftmax(dim=-1)

def train_step(...):
    prob = log_softmax(model_output)
    loss = 0
    for i in range(sequence_size):
        loss += itf_loss(prob[i], target[i])
```

4 Experiments

We experimentally compared the diversity of the dialogue generation of our ITF loss model and previous methods using three dialogue datasets in a couple of different domains and languages.

4.1 Training Details

We chose $\lambda = 0.4$ for all the ITF loss models based on the discussion below in Section 4.5. In the decoder, we applied a repetition suppressor with $\lambda = 1$ in all the models to suppress the repetitive generation of identical phrases for improving the quality. Details are discussed in Section 4.7.

In all the models, we set four layers in both the encoder and the decoder, 256 hidden units, an embedding size of 256, a maximum sequence size of 28, a mini-batch size of 32 and trained them with the Adam Optimizer at a learning rate of 3e-4. We tokenized the sentences by a subword model with a 32k vocabulary using Sentencepiece (Kudo 2018).

4.2 Baselines

We compared our loss model to some competitive models.

Seq2Seq An encoder-decoder (a.k.a., sequence-to-sequence) network has been applied to many generation-based dialogue systems (Shang, Lu, and Li 2015; Vinyals and Le 2015; Sordoni et al. 2015). We used it with the bidirectional multi-layered LSTM encoder and the multi-layered LSTM decoder, both of which have residual connections around each layer. The bidirectional LSTM encoder compresses well the feature representation of the whole source sentence, and the residual connection helps train the deep neural networks.

Seq2Seq + Attn An attention mechanism improved the performance and the diversity by referring to encoded memory (Zhou et al. 2017; Shao et al. 2017). In the above basic Seq2Seq, as the decoding process continues, the constraints from the source sentence often weaken, and then the decoding depends on the generated tokens like in a language model. Since the attention mechanism refers to the feature representation of the source sentence at each time step, it helps avoid language model-like generation and increases diversity. We use the encoder-decoder, which controls the decoder by Scaled Dot-Product Attention (Vaswani et al. 2017).

Seq2Seq + MMI Based on MMI-antiLM inference (Li et al. 2016a), the Maximum Mutual Information objective function is defined:

$$\hat{T} = \arg\max \left\{ \log p(T|S) - \lambda \log p(T) \right\}, \quad (4)$$

where $\log p(T|S)$ is the conditional log-likelihood of a generated sentence given a source sentence and $\log p(T)$ is the unconditional log-likelihood of the generated sentence as a language model. By subtracting a language model term, MMI-antiLM suppresses language model-like generation. Note that the diversity does not improve when MMI-antiLM is used during the training. As described by Li et al. (2016a), we used MLE during the training and MMI-antiLM during the evaluation. In practice, MMI-antiLM generates token y:

$$y = \arg\max \left\{ \log \text{softmax}(x - \lambda u) \right\}, \quad (5)$$

where $x \in \mathbb{R}^{|V|}$ is the output of the projection layer using the encoder-decoder given a source sentence and $u \in \mathbb{R}^{|V|}$ is the output of the projection layer using only the decoder (i.e., initial value of the LSTM's hidden state is set to zeros). Other formulations, such as $\log \text{softmax}(x) - \log \text{softmax}(\lambda u)$ and $\log(\text{softmax}(x) - \text{softmax}(\lambda u))$, did not work well in our preliminary experiment. Coefficient λ is the degree of the anti-language model. We chose $\lambda = 0.8$ for all the datasets and only applied MMI-antiLM to the first five time steps of the decoder (i.e., $\gamma = 5$).

MemN2N In dialogue generation, models can acquire contextual consistency by referring to multi-turn utterances as a dialogue history. The Memory Network (MemN2N) and the Hierarchical Recurrent Encoder-Decoder (HRED) are typical ways to encode multiple utterances (Sukhbaatar et al. 2015; Miller et al. 2016; Serban et al. 2016; 2017b). We use the former, which encodes the dialogue histories of multi-turns, stores them in memory slots, and extracts contex-

tual information by attention. We generated a sentence vector from a tokens matrix with a bidirectional multi-layered LSTM instead of summation with positional encoding. We always applied temporal encoding.

4.3 Evaluation Details

We used BLEU to measure the quality of the generated sentences and DIST to measure the diversity. The following are the details of each metric.

BLEU BLEU-n calculates the percentage of n-gram matching between all of the generated sentences and all of the reference sentences (Papineni et al. 2002). We calculated the corpus-level BLEU-1 and BLEU-2 scores that measure the degree of unigram and up to bigram matching. We also applied a brevity penalty that incorporated recall and a smoothing method that added 0.1 counts to precision with 0 counts.

Some dialogue generation studies obtained BLEU-4 scores, but in our experiments the BLEU-4 scores were very low, typically less than 2. Because there are an enormous number of generation candidates, higher-order n-grams are hardly matched in the reference, and scores slide up and down depending on the initializing model and the sampling differences of the mini-batches. Therefore, the corresponding BLEU-4 scores become more unstable.

DIST DIST-n calculates the percentage of the distinct n-grams in all the n-grams of the generated responses proposed by (Li et al. 2016a). We calculated the DIST-1 and DIST-2 scores that measure the degree of the unigram and bigram diversity.

We tokenized with the TweetTokenizer in the NLTK to calculate the BLEU and DIST scores for the word sequences (not subword sequences). Note that for the Japanese Twitter replies, we tokenized with SentencePiece to directly calculate the BLEU and DIST scores for the subword sequences because no Japanese tokenizer was suitable for tweet data. We removed such symbols as Padding, Unknown, Start, and End from all the sentences during the evaluations. Because a beam search maximizes the likelihood of the whole sentence and causes low diversity, the decoders of all the models generate tokens by a greedy search.

4.4 Results

OpenSubtitles We extracted dialogue data from the OpenSubtitles2018 corpus (Lison and Tiedemann 2016). This corpus has multiple subtitles for the same movie, but we used only one subtitle per movie to avoid an imbalanced training set. In this corpus, we obtained the start and end times of each turn of the subtitles. Each episode was configured as continuous turns with the interval from the end time of a turn to the start time of the next turn within five seconds. As a result, the OpenSubtitles training set consists of 5M turns and 0.4M episodes (i.e., 4.6M examples). Since all the episodes have multi-turns, we can use the memory network to consider the dialogue history. The validation and test sets have 10k examples respectively.

Table 2: Results of English OpenSubtitles dialogue. BLEU-1/2 calculates percentage of unigram/up to bigram matching. DIST-1/2 calculates percentage of distinct unigram/bigram in generated responses. Reports of previous works use a different number of examples in training/test sets.

Method	BLEU-1/2	DIST-1/2	Length
Reference	100/100	8.68/44.4	7.21
MMI (2016a)	-	1.84/6.6	-
RL (2016b)	-	1.7/4.1	-
DP-GAN (2018)	-	2.39/11.1	-
Seq2Seq	13.3/2.95	1.43/4.79	5.78
MemN2N	**13.6**/3.11	1.80/7.20	5.76
Seq2Seq + Attn	13.3/**3.67**	4.02/14.1	5.56
Seq2Seq + MMI	12.2/2.53	6.54/**25.9**	5.32
Seq2Seq + ITF	12.9/2.70	**7.56**/21.6	**6.07**

Table 3: Results of English and Japanese Twitter replies

Method	English BLEU-1/2	DIST-1/2	Length
Reference	100/100	10.2/57.3	13.7
Seq2Seq	**10.6/3.25**	1.25/5.66	**11.3**
Seq2Seq + MMI	7.12/1.66	6.06/**33.0**	9.24
Seq2Seq + ITF	7.50/2.14	**7.67**/26.3	10.3

Method	Japanese BLEU-1/2	DIST-1/2	Length
Reference	100/100	16.2/71.0	10.9
Seq2Seq	10.7/2.63	7.98/26.3	6.11
MemN2N	**13.8/3.84**	7.83/29.8	7.12
MemN2N + MMI	12.6/2.69	14.5/**55.1**	7.27
MemN2N + ITF	12.8/3.03	**16.8**/54.3	**8.27**

Table 2 shows that our Seq2Seq trained with ITF loss establishes a state-of-the-art DIST-1 of 7.56 while maintaining a good BLEU score. Regarding the relative improvement of DIST-1 from the baseline Seq2Seq, MMI-antiLM (Li et al. 2016a) reported 228%, RL (Li et al. 2016b) reported 174%, and DP-GAN (Xu et al. 2018) reported 25%, but our ITF loss model achieved 429%. Seq2Seq with MMI inference increased DIST, but slightly decreased BLEU. Seq2Seq with Attention increased BLEU-2 and DIST. MemN2N achieved the highest BLEU-1 of 13.6, but its DIST improvement was small.

Twitter We collected datasets of both English and Japanese Twitter replies. We excluded self-replied dialogues, bot-to-bot dialogues, and extremely long dialogues from these data. The English Twitter training set consists of 5M turns and 2.5M episodes (i.e., 2.5M examples). All episodes have two turns. The Japanese Twitter training set consists of 4.5M turns and 0.7M episodes (i.e., 3.8M examples). All of the episodes have multi-turns. On both the English and Japanese datasets, the validation and test sets have
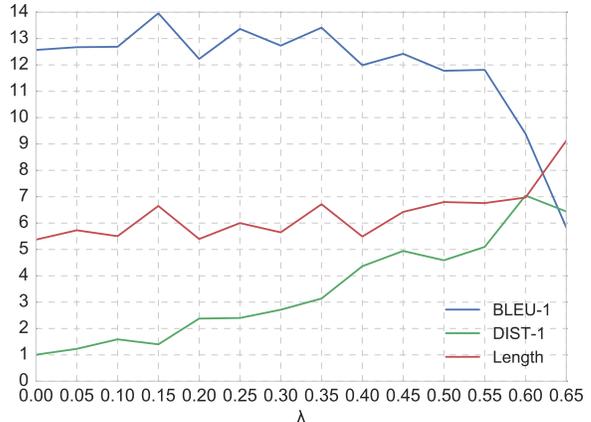

Figure 1: Comparison of automatic evaluation scores for each λ in ITF loss on OpenSubtitles dialogue. Number of turns in training set is 500k, which is $1/10$ compared with Section 4.4 and the number of subwords is 16k.

Table 4: Comparison of automatic evaluation scores for each inference method. ITF infer is different from ITF loss.

Inference	λ	BLEU-1	DIST-1
Noisy infer.	1.4	12.5	2.81
MMI infer.	0.7	12.5	4.76
ITF infer.	0.09	12.5	4.85

10k examples respectively.

Table 3 shows that on both the English and Japanese datasets, our ITF loss model outperforms the MMI inference model on both BLEU-1 and DIST-1. In particular, on the Japanese dataset, our loss model achieved a DIST-1 score of 16.8 compared to a ground truth of 16.2.

4.5 Selection of λ in ITF Loss

We investigated the optimal value of hyperparameter λ in Eq. 3 through which the ITF loss model yields high diversity while maintaining good quality. We used a set of $\lambda = \{0, 0.05, \dots, 0.6, 0.65\}$ and trained Seq2Seq on an OpenSubtitles dialogue dataset that consists of 500k turns.

Figure 1 shows the results. The generated sentences have a sufficiently high DIST-1 while maintaining a high BLEU-1 using λ around 0.4.

4.6 ITF Inference in MLE Model

We introduce another inference version of ITF loss, which applies the concept of inverse token frequency to the model trained with MLE during the evaluation. It resembles the use of MMI inference (Li et al. 2016a). One advantage is that it is unnecessary to re-run the training when we use different λ values, compared to using ITF loss during the training. ITF inference generates token y:

$$y = \operatorname{argmax} \{\log \operatorname{softmax}(w \odot x)\}, \qquad (6)$$

Table 5: Examples of repetitive generation and its suppression

do nothing
i'm sorry to hear that. i'm sorry to hear that.
i'm not sure i'm a cop. i'm not a cop. i'm not a cop.
suppress repetition ($\lambda = 1$)
i'm sorry to hear that. hope you enjoyed it!
i'm not sure how you can do that.

Table 6: Percentage of generated sentences containing identical tokens in all generated sentences. $\lambda = 0$ means that no repetition suppressor was used.

Dataset	Repetition		
λ	0	0.5	1
OpenSubtitles	6.22	1.08	**0.93**
English Twitter	64.4	25.3	**21.8**
Japanese Twitter	45.1	16.4	**6.66**

where $x \in \mathbb{R}^{|V|}$ is the output of the projection layer, $w \in \mathbb{R}^{|V|}$ is the weight (i.e., the vector version of Eq. 3), and $\odot$ is the element-wise product.

We also introduce a noisy inference to prove that the ITF and MMI inferences have more meaning than just noise injection:

$$y = \operatorname{argmax} \left\{ \log \operatorname{softmax}(x + \lambda \operatorname{noise}) \right\}, \quad (7)$$

where $\operatorname{noise} \in \mathbb{R}^{|V|}$ is the sampling from standard normal distribution $\mathcal{N}(0, 1)$.

Table 4 shows that the performance of our ITF inference is close to MMI inference, and both are superior to noisy inference. We chose each λ to be equivalent BLEU scores.

4.7 Suppression of Repetitive Generation

In our preliminary experiment, the decoder generated repetitive phrases (Table 5), which gravely decreased the quality of the generated responses. This problem can be avoided by suppressing the regeneration of the already generated tokens during the decoding process. We defined a repetition suppressor:

$$\operatorname{suppressor}(x_k) = \frac{1}{\{1 + \operatorname{count}(\operatorname{token}_k)\}^\lambda} x_k, \quad (8)$$

where x_k is the k-th element of $x \in \mathbb{R}^{|V|}$, which is the output of the projection layer and $\operatorname{count}(\operatorname{token}_k)$ is the number of times token_k was generated in previous time steps during the decoding process.

We calculated the percentage of the generated sentences containing the same token in all the generated sentences. Table 6 shows that the repetition suppressor avoids the regeneration of the previously generated tokens in all the datasets.

5 Discussion

Methods and results answer the following two questions about ITF loss.

1) Does ITF loss potentially generate grammatically incorrect sentences more often?

Our generated responses in the OpenSubtitles dialogue (Table 7) and the Twitter replies (Table 8) show that the ITF loss model yields high diversity and quality that exceed or equal the result of the Maximum Mutual Information (MMI) inference model. We found no serious grammatical problems in the generated sentences when it is less than $\lambda = 0.7$.

2) Does ITF loss modify the underlying distribution of tokens when generating a response?

The goals of SCE and ITF loss are the same: getting the correct distribution of tokens. ITF loss is not directly used to modify the distribution of tokens. It only applies to expected errors (a.k.a., SCE loss) from the distribution of tokens to avoid the overvaluation of frequent tokens. In other words, scale w in Eq. 3) just controls its contributions of the errors and ITF loss controls the gradients on individual tokens by frequency-based scaling. This speeds up the learning on rare tokens and slows it down on frequent tokens.

6 Conclusion

We focused on the low diversity problem and confirmed that unigram diversity scores significantly improve by applying Inverse Token Frequency loss. Future work will investigate Inverse N-gram Frequency (INF) loss, which is a general type of ITF loss that only considers unigram frequency. Since BLEU is not well suited to evaluate dialogue responses, we are planning to conduct human evaluations.

References

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; and Bengio, S. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 10–21.

Cao, K., and Clark, S. 2017. Latent variable dialogue models and their diversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2, Short Papers)*, volume 2, 182–187.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kudo, T. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Table 7: Generated responses on OpenSubtitles dialogue. SRC and TGT denote source and target sentences.

SRC	**Does he know what's going on?**
TGT	**He knows he's leaving.**
MLE	No.
MMI	No.
ITF	He's got a lot of trouble.

SRC	**I'm not interested in talking to you.**
SRC	**I'm leaving.**
TGT	**Have the trial by yourselves.**
MLE	I'm going to the bathroom.
MMI	Okay.
ITF	I'll be back tomorrow morning.

SRC	**And you don't know Batman.**
SRC	**He's not letting this lunatic just run wild.**
TGT	**He can't catch him either.**
MLE	You're not going to be a mutant.
MMI	Then why he's not here?
ITF	He's a hero.

SRC	**My client, Turatello, is coming.**
SRC	**There'll be chaos.**
SRC	**They did a great job on your ass.**
TGT	**Get well soon.**
MLE	I'm sure you have a good friend.
MMI	But you have no idea how much I'm going to get away.
ITF	I was hoping to be honest, but you're not interested in the first time.

SRC	**It's not loaded, look.**
SRC	**See?**
SRC	**We just need some help, that's all.**
TGT	**Our friends, they're trapped.**
MLE	I'm sorry.
MMI	Fine, syringes.
ITF	What do you want?

SRC	**Sorry to leave you in the line.**
SRC	**At the end of it, ok.**
SRC	**They all cheat in Italy.**
TGT	**Off you go.**
MLE	What?
MMI	They sell boats?
ITF	They're not idiots who is in Italy, but we can make a lot of trouble.

Table 8: Generated responses on English Twitter replies. SRC and TGT denote source and target sentences.

SRC	**i hope you had a wonderful trip! can that flight please pick me up to go to iceland? what a beautiful place!**
TGT	**place it on your short list of "to-do's.". it's a beautiful country.**
MLE	i'll be back in december!
MMI	poutine flight iceland has a great time.
ITF	yes! iceland is awesome :)

SRC	**i don't think it'll happen, but apple does have the money..**
TGT	**i suspect the design teams are fundamentally incompatible. mclaren much more radical and more performance oriented.**
MLE	i'm not sure if it is, but the fact that the android phone is a bad thing.
MMI	true apple products are the best.
ITF	i think apple is working on android aosp, pixel phones are very good.

SRC	**12 gb ram at 384 gb/sec (gddr5x).. if this is true than damnn!.. you want**
TGT	**eurogamer also said that it speculates 384 and 12 gb of ram by placement of ram in original video**
MLE	i'm not sure if it is worth it.
MMI	mwr gpu is the best.
ITF	rambo is a newer one with chromebook7 connector, laptop router, hdmi cables.

SRC	**10 reasons you need to exercise with your bodyweight**
TGT	**yep, planks, side planks, etc are my go to!**
MLE	i'm not a good person, but it is hard to get through.
MMI	metabolism thighs lotion depending on how you feel
ITF	10 yrs later: 1.15 pounds of weight rate increases so much easier than a lot.

SRC	**im going back to dragon ball z, while yall talking about pikachu, ill be talking about goku**
TGT	**which goku**
MLE	i'm not even gonna be able to play that game
MMI	vegeta and goku literally
ITF	i knoww, goku rap eminems remix

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 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*, 110–119.

Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1192–1202.

Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2157–2169.

Lison, P., and Tiedemann, J. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation.*

Miller, A.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1400–1409.

Olabiyi, O.; Salimov, A.; Khazane, A.; and Mueller, E. 2018. Multi-turn dialogue response generation in an adversarial learning framework. *arXiv preprint arXiv:1805.11752.*

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence.*

Serban, I. V.; Sankar, C.; Germain, M.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Pieper, M.; Chandar, S.; Ke, N. R.; et al. 2017a. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349.*

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence.*

Shang, L.; Lu, Z.; and Li, H. 2015. 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)*, volume 1, 1577–1586.

Shao, Y.; Gouws, S.; Britz, D.; Goldie, A.; Strope, B.; and Kurzweil, R. 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*, 2210–2219.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 196–205.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, 2440–2448.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869.*

Wen, T.-H.; Gašic, M.; Kim, D.; Mrkšic, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015a. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 275.

Wen, T.-H.; Gasic, M.; Mrksic, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015b. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1711–1721.

Xu, J.; Sun, X.; Ren, X.; Lin, J.; Wei, B.; and Li, W. 2018. Dp-gan: Diversity-promoting generative adversarial network for generating informative and diversified text. *arXiv preprint arXiv:1802.01345.*

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence.*

Zhao, T.; Zhao, R.; and Eskenazi, M. 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, 654–664.

Zhou, G.; Luo, P.; Cao, R.; Lin, F.; Chen, B.; and He, Q. 2017. Mechanism-aware neural machine for dialogue response generation. In *Thirty-First AAAI Conference on Artificial Intelligence.*