Diverse dialogue generation with context dependent dynamic loss function

Ayaka Ueyama, Yoshinobu Kano
Faculty of Informatics, Shizuoka University, Japan
aueyama@kanolab.net, kano@inf.shizuoka.ac.jp

Abstract

Dialogue systems using deep learning have achieved generation of fluent response sentences to user utterances. Nevertheless, they tend to produce responses that are not diverse and which are less context-dependent. To address these shortcomings, we propose a new loss function, an Inverse N-gram Frequency (INF) loss, which incorporates contextual fluency and diversity at the same time by a simple formula. Our INF loss can adjust its loss dynamically by a weight using the inverse frequency of the tokens' n-gram applied to Softmax Cross-Entropy loss, so that rare tokens appear more likely while retaining the fluency of the generated sentences. We trained Transformer using English and Japanese Twitter replies as single-turn dialogues using different loss functions. Our INF loss model outperformed the baselines of SCE loss and ITF loss models in automatic evaluations such as DIST-N and ROUGE, and also achieved higher scores on our human evaluations of coherence and richness.

1 Introduction

Recently, many reports have described studies using deep learning for dialogue systems that have achieved good performance. They can generate fluent sentences based on a user's utterances (Vinyals and Le, 2015; Shang et al., 2015; Serban et al., 2016). Nevertheless, such neural dialogue systems tend to generate phrases such as "Yes" and "I do not know" frequently in non-task-oriented dialog systems, referred to as the low diversity issue and the generic response issue. After training by a loss function of similarity with gold standard reference sentences, frequent phrases are more likely to be assigned a large occurrence probability than rare phrases are.

Nakamura et al. (2018) proposed an Inverse Token Frequency (ITF) loss, which multiplies the Softmax Cross-Entropy (SCE) loss by weights based on the inverse of the frequency of tokens. This ITF loss incorporates the frequency distribution of token classes so that rare tokens become more likely to appear.

However, sentence diversity is based not only on individual tokens, but also on the token sequence. We are able to compute weights for loss functions dynamically, depending on the context, while retaining the fluency of generated sentences. We propose such a loss function, Inverse N-gram Frequency (INF) loss, which uses the inverse of the frequency of the n-gram of the tokens, rather than the token frequency. We built a neural dialogue system trained by INF loss using huge amounts of dialogue data extracted from Twitter. After comparing models using the SCE loss, the ITF loss, and the INF loss, we evaluated their diversity and fluency. Results show that our proposed INF loss model outperformed the SCE loss and ITF loss models for most automatic assessment measures such as DIST-N (Li et al., 2016) and ROUGE (Lin, 2004). Our INF loss model also achieved higher scores on our human evaluations of coherence and richness.

2 Related Work

The diversity of neural dialogue generation has been studied actively. Li et al. (2016) first addressed this problem using Maximum Mutual Information (MMI) as the objective function of the neural model. Takayama and Arase (2019) used Positive Pointwise Mutual Information (PPMI) to identify keywords

Proceedings of the 28th International Conference on Computational Linguistics, pages 4123–4127
Barcelona, Spain (Online), December 8-13, 2020

in the dialogue corpus that were likely to appear both in response utterances and their input utterances. Xing et al. (2017) proposed a model that uses topic words extracted from conversations to simulate human prior knowledge, generating informative and interesting responses. In addition, Variational AutoEncoder (VAE) and Generative Adversarial Network (GAN), which were proposed originally for image generation, have also been applied to text and dialogue generation (Kingma and Welling, 2014; Bowman et al., 2016; Xu et al., 2018). Although GAN helps to reduce response text ambiguity, their primary purpose was not diversity. Zhang et al. (2018) proposed and demonstrated the effectiveness of Adversarial Information Maximization (AIM) as a new method for generating informative and diverse conversational responses. Their work also resolved instability that arose when training the GAN model.

3 Loss Functions

3.1 Softmax Cross-Entropy Loss

The SCE loss, which is often used to train a sequence-to-sequence (Seq2Seq) model (Sutskever et al., 2014), is expressed as $L_{SCE} = -log\ \{softmax_c\}$, where $softmax_c = \frac{e^{d_c}}{\sum_k^{|V|} e^{d_k}}$. Therein, V represents the lexicon; d_k denotes the k-th element of the output $d \in \mathbb{R}^{|V|}$.

3.2 Inverse Token Frequency Loss

Nakamura et al. (2018) defined Inverse Token Frequency (ITF) loss as shown below.

$$L_{ITF} = w_c L_{SCE}, \quad w_c = \frac{1}{freq(token_C)^\lambda} \tag{1}$$

Therein, where w_c represents an element of the weight $w \in \mathbb{R}^{|V|}$ of the token class c, $token_C$ is a token of token class c, $freq(token_C)$ stands for a function that counts $token_C$ in the training data, λ denotes a hyper-parameter that adjusts the frequency. The ITF loss at λ=0 is equivalent to SCE loss.

3.3 Inverse N-gram Frequency Loss

We propose our Inverse N-gram Frequency (INF) loss as a new loss function that replaces the ITF loss. Our INF loss is defined as presented below.

$$L_{INF} = w_c L_{SCE}, \quad w_c = \frac{1}{freq\{ngram_C(n)\}^\lambda} \tag{2}$$

Therein, $ngram_C(n)$ represents n consecutive ($n \in \mathbb{N}, n \geqq 2$) tokens in the training data, where c is the first token of $ngram_C(n)$. Therefore, w_c is expected to differ considerably depending on the context.

Special symbols such as <BOS> (beginning of sentence) and <EOS> (end of sentence) are treated similarly to other ordinary tokens. However, because <BOS> and <EOS> occur in every sentence, we set the weight of the loss function as very small value for these special characters.

A special symbol, <NORM>, was added n- 1 times to the beginning of the sentence and to the end of the sentence respectively to represent padding.

4 Experiments

The SCE model and the ITF model were used as baseline models. Experiment settings are the same among these models and our proposed INF model, except for loss functions.

4.1 Dataset

Japanese and English Twitter conversations were extracted from Twitter replies, adjacent tweets as pairs of [utterance, response] to construct a single-turn dialogue dataset of one million dialogue pairs for each language. SentencePiece (Kudo and Richardson, 2018) was trained using a dataset with a vocabulary of 32,000 for both Japanese and English data. We then used these SentencePiece models to tokenize the training set into subwords. Each of the verification set and the test set consists of 1024 pairs.

4.2 Model Setting

Both the encoder and decoder are six-layer Transformer (Vaswani et al., 2017), the number of heads of Multi-Head Attention is 8, the token embedding dimension is 512, and the ratio of Dropout is 0.1. Adam

model	λ	perplexity	BLEU-1	BLEU-2	DIST-1	DIST-2	ROUGE-1	ROUGE-2	length
SCE	-	60.572*	15.19	0.049	1.376*	7.949*	3.781	0.421	8.013*
ITF	0.2	16.512*	15.32	0.016	1.303*	6.592*	3.775	0.357*	7.513*
	0.4	9.401*	**15.78**	0.033	0.524*	2.594*	2.916*	0.305*	7.737*
	0.6	7.511*	15.15	0.034	0.245*	1.309*	2.308*	0.181*	6.002*
	0.8	7.341*	13.71	**0.082**	0.155*	0.697*	1.873*	0.096*	6.350*
INF	0.2	11.045	13.97	0.049	**3.265**	**15.285**	**3.877**	**0.445**	10.253
	0.4	5.936	12.35	0.049	1.401	5.661	2.482	0.295	**10.521**
	0.6	4.413	12.12	0.033	2.001	6.981	2.633	0.234	10.114
	0.8	**3.890**	12.32	0.033	2.039	6.681	1.348	0.085	9.998

Table 1: Metric-based evaluation results in Japanese data. (%)

model	λ	perplexity	BLEU-1	BLEU-2	DIST-1	DIST-2	ROUGE-1	ROUGE-2	length
SCE	-	107.582*	6.301	0.011	0.068*	0.576*	0.762*	0.156	49.335
ITF	0.2	23.475*	6.141	0.012	0.117*	1.138*	2.006	0.141	**55.884**
	0.4	10.227*	**6.583**	0.005	0.097*	1.323*	1.443*	0.142	35.889
	0.6	35.383*	5.455	**0.016**	0.085*	1.344*	1.325*	0.085*	47.063
	0.8	34.909*	5.517	**0.016**	0.068*	1.296*	0.653*	0.061*	34.266
INF	0.2	22.464	6.516	0.011	0.118	1.377	**2.014**	**0.171**	54.451
	0.4	6.686	6.357	0.005	**0.169**	**1.473**	1.439	0.128	40.339
	0.6	4.534	6.414	0.013	0.092	0.974	0.616	0.105	44.106
	0.8	**3.796**	6.508	0.011	0.118	0.871	0.954	0.041	45.716

Table 2: Metric-based evaluation results in English data. (%)

(Kingma and Ba, 2015) was used as the optimization method for parameters during training. The learning rate of Adam was set to 0.001. Hyperparameters λ, which adjust the frequency of ITF model and INF model, were set as 0.2, 0.4, 0.6, or 0.8. The INF model uses bi-gram as its n-gram function.

5 Results

5.1 Metric-Based Evaluation

Table 1 and Table 2 respectively present evaluation results for Japanese and English datasets in perplexity, BLEU (Papineni et al., 2002), DIST-N (Li et al., 2016), ROUGE (Lin, 2004), also showing length, which is an average number of tokens generated in a sentence. * in these tables indicate significant differences between baseline models and INF model for each evaluation metric ($p<0.05$). BLEU and ROUGE were used to assess the quality of the generated sentences, whereas DIST-N was used to calculate the proportion of different n-grams among the n-grams included in the generated sentences, and therefore to assess the diversity of the generated sentences.

Regarding the Japanese dataset, the best perplexity value was obtained using the INF model when λ = 0.8. Results show that INF and ITF performed better than SCE. In fact, ITF yielded the best scores for both unigram and bigram BLEU scores. INF with λ = 0.2 yielded the best scores for DIST-N and ROUGE. The length of the generated sentences was greater in the order INF > SCE > ITF. Overall, the INF model at λ = 0.2 is superior in the metric-based evaluation. The English dataset yielded similar results: the INF model was superior.

5.2 Human Evaluation

model	λ	Coherence	Richness
SCE	-	2.132/2.032	2.163/2.007
ITF	0.4	2.071/2.101	2.002/2.094
INF	0.2	**1.794**/**1.866**	**1.835**/**1.899**

Table 3: Human evaluation. (Japanese/English)

Table 3 shows our human evaluation results obtained using Japanese and English models. For each dataset, 100 dialogue tweet pairs were presented to five human evaluators. Responses for each pair include responses of four patterns: generated responses of our three models and the original tweet reply. Evaluators ranked the responses generated by the respective models in terms of coherence and richness, with scores of 1, 2, and 3. Richness was defined as whether evaluators would

ID		Japanese	ID		English
1J	SRC	The theory that he's more active than a star 5.	1E	SRC	You can get Supernova free now! Who's downloading? We'd love to see what you can create!
	TGT	I think there is. You are especially welcome if you are a stray!		TGT	I would but You need to launch a web version ;) PC user here?
	SCE	Seriously?		SCE	I'm not sure if I'm not going to get into it. I'm just saying, "I'm gonna be the next."
	ITF	Really ! lol		ITF	I'm going to rewatch soon.
	INF	Well, I think this star 4 is stronger than star 5.		INF	I don't like the songs, but I've listened to it.
2J	SRC	I want to collect all the DVDs.	2E	SRC	Honestly it grew on me later but god he ruined it.
	TGT	It's in the Mozuku mansion.		TGT	nooo he didn't ruin it its still a bop come on?
	SCE	Are you buying <unk>4?		SCE	Is this the same guy that has to be in the same position?
	ITF	It's amazing how everything is <unk>.		ITF	I'm sorry sis but I'm gonna be honest.
	INF	They were all 980 yen.		INF	I'm a big fan of his face, but I've never listened to the songs.

Table 4: Example responses generated from different models.

be willing to talk to that systems in the future. A smaller score denotes a better evaluation rank. We used a model setting with the best metric-based evaluation score for each loss function model, i.e. λ=0.4 for the ITF model and λ=0.2 for the INF model. Table 3 presents the average human evaluation rankings. These results demonstrate that INF was best in terms of both coherence and richness, followed by ITF and SCE.

6 Discussion

Results show that INF produced superior perplexity values compared to SCE and ITF. Therefore, INF produces more fluent sentences than the baselines. BLEU and ROUGE often showed good results in SCE and ITF. BLEU calculates n-gram precision between generated and target sentences, which does not incorporate recall. Therefore, SCE and ITF can obtain better performance because their generated sentence lengths were shorter. They are not directly comparable with the INF of the longer sentences.

Also, INF presents excellent values for the diversity in both DIST-1 and DIST-2. Particularly, the INF of DIST-2 for the Japanese model and the Japanese dataset was 15.285, which is much higher than SCE and ITF. This result demonstrates that INF can generate responses that are more diverse, which also incorporate sequential information of tokens.

Table 4 presents example dialogue pairs, where SRC is the input sentence to the model and TGT is the target sentence, i.e. an original reply tweet. The λ values of ITF and INF are the same as the model used for human evaluation. Regarding the example ID1J, INF generated a rich contextual response, whereas the responses generated by SCE and ITF were less context-dependent. In the ID2J dialogue, all three models generated responses that make sense as replies to the input, but they generated different tokens. In such cases, automatic evaluation using a single reference output might not be meaningful. In the ID1E dialogue, SCE generated a sentence that does not make sense. INF generated a meaningful sentence related to a song probably because "SUPERNOVA" was interpreted as a Korean dance group, although it differs from the human reference response. In the ID2E dialogue, responses that are more relevant are potentially available because its input sentence is ambiguous; SCE and ITF were unable to generate any meaningful conversation, although INF was able to generate them.

Overall, all three models generated semantically relevant responses, but INF generates a wider variety of responses with longer sentences.

7 Conclusion

We proposed a weighted INF loss based on the inverse of the frequency of token n-grams, which can generate diverse responses while retaining fluency. Comparison to baseline methods revealed that our automatic and human evaluation scores obtained using our proposed method generated more diverse responses with improved contextual consistency. Future works are expected to include a tri-gram version of our INF model and the use of deep learning models other than Transformer.

Acknowledgements

This research is supported by KAKENHI (18K18504) and JST CREST (JPMJCR1684, JPMJCR19F4).

Reference

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. 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.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. *Computing Research Repository*, arXiv: 1312.6114. Version 10.

Diederik P. Kingma, and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *In Proceedings of the Third International Conference on Learning Representations*, pages 3351–3357.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 74–81.

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. 2018. Another diversity promoting objective function for neural dialogue generation. *In Proceedings of the second AAAI Workshop on Reasoning and Learning for Human-Machine Dialogues*.

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 of the Association for Computational Linguistics*, pages 311–318.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. *In Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1577–1586.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *In Proceedings of the 28th Conference on Neural Information Processing Systems*, pages 3104–3112.

Junya Takayama and Yuki Arase. 2019. Relevant and Informative Response Generation using Pointwise Mutual Information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of International Conference on Machine Learning, Deep Learning Workshop*.

Chen Xing, Wei Chung Wu, Yu Ping Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. *In Proceedings of the 31st AAAI Conference on Artificial Intelligence,* pages 3351–3357.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31:1810–1820.