Efficient Diverse Response Generation in Attention-based Neural Conversational Model with Maximum Mutual Information

Yuki Kisida*, Tsuneo Kato*, Yanan Wang †, Jianming Wu † and Gen Hattori †
* Doshisha University, Kyoto, Japan
E-mail: ctwd0125@mail4.doshisha.ac.jp
† KDDI Research, Inc., Saitama, Japan

***Abstract*—Diversity of generated responses is important for a data-driven neural conversational model (NCM) for non-task-oriented conversation. A criterion of maximum mutual information (MMI) and generating N-best outputs are both effective ways to increase the diversity. Generally, a beam search is used for generating N-best outputs. However, the beam search is likely to produce similar outputs in the N-best results. We propose a simple and efficient N-best search, namely N-greedy search, for an encoder-decoder recurrent neural network (RNN) with an attention mechanism. We built an NCM with a fictive chitchat corpus and generated responses based on the MMI criterion and N-greedy search. All of four objective indices of diversity showed increases, and a subjective evaluation clearly showed a reduction in the number of dull responses.**

I. Introduction

In accordance with today's prevalence of AI assistants and chatbots, non-task-oriented response generation techniques are gaining attention. This type of conversation is important for an agent to build rapport with a user through continuing coherent and diverse responses to idle talks.

Applications of sequence-to-sequence neural network models to the conversational domain have shown fluent response generation [1], [2], [3]. However, such generative models trained on the basis of the criterion of minimizing cross-entropy error have suffered from generating dull responses such as "I think so, too" and "I don't know."

To address this issue, Li et al. proposed diverse response generation on the basis of a criterion of MMI with an input sentence [4]. The concept of MMI was introduced to model training and implemented with an incremental reinforcement learning (RL) algorithm called MIXER [5]. Reinforcement learning enabled introduction of various types of rewards such as ease of answering and semantic coherence [6]. MMI was further incorporated into a generative adversarial network (GAN) [7].

Outputting N-best results is also effective for diverse response generation. The N-best results leave the possibility of choosing one on the basis of different criteria. For example, the N-best results were compared and re-ranked based on event causality relations [8]. The N-best results are also effective for robust model training. To circumvent "exposure bias", a frequnecy gap of words that appear in training and in decoding, sequence level training considering various hypotheses with an optimized beam width showed effectiveness in multiple tasks including machine translation [9]. However, a beam search tends to produce similar sentences that diverged from a common partial sentences and different at the end of the sentences. The N-best results are desired to be efficient in terms of diversity. In other words, a variety of fluent sentences are desired to be contained in N-best results. Regarding this issue, iterative beam search which runs multiple iterations of beam search with excluding any previously explored space was proposed [10]. Recently, a simple but effective variation of top k sampling, "nucleus sampling" was proposed to generate diverse but coherent responses [11].

We developed an encoder-decoder NCM with an attention mechanism using a large-scale fictive chitchat corpus between two female personae. To obtain diverse responses efficiently with N-best outputs, we propose a simplified N-best search, "N-greedy search". In addition to the MMI criterion that reduced dull responses substantially, N-greedy search further improved the diversity of the responses with a simple implementation.

II. Fictive chitchat corpus

We collected a fictive chitchat corpus between two Japanese female personae in their twenties through crowd sourcing. To let crowd workers share common images of the personae, we set 80 attributes from basic ones, such as name, birthday, family, friends, and job, to detailed ones such as personality, various favorites and dislikes, current mood and today's time of waking up. One of the two initiates chitchats in 50 different ways, and the ladies exchange responses alternately up to 10 times. Crowd workers were asked to compose three likely responses to each input from the interlocutor with the personae in mind. Thus, each of 50 initial sentences formed a tree of triply-branched alternate responses with a depth of 10.

A total of 1.68 million responses were composed by 200 crowd workers in a time span of 10 months. We assessed the quality of the composition in two aspects: the rate of reproduction among three responses to an input along with the rate of inconsistency between responses and the persona settings by random sampling. The quality was considered to be good enough with a 7% reproduction rate and a 5% inconsistency rate.

III. Neural Conversational Model

A. Base model

The base model is an encoder-decoder recurrent gated unit (GRU) [12] neural network with an attention mechanism [13]. The encoder accepts the previous input from the interlocutor, and the decoder generates a response with an attention to the encoder's output vectors corresponding to tokens of the input from the interlocutor. The encoder is composed of an embedding layer and a GRU layer. The decoder is composed of an embedding layer, an attention layer, a GRU layer, and a fully-connected layer with a softmax function.

Let the previous input from the interlocutor and its response represented by $S = \{s_1, \ldots, s_{N_S}\}$ and $T = \{t_1, \ldots, t_{N_T}\}$, where s and t denote indices of source (input) and target (output) tokens. The indices are common to inputs and outputs with a vocabulary size V. In the encoder, a GRU accepts a concatenation of an embedding vector of an input token $\boldsymbol{x}_{s_j}$ and a hidden state at the previous token $\boldsymbol{h}^e(j-1)$, and updates the hidden state with an output of an encode vector $\boldsymbol{e}_j$.

$$\boldsymbol{x}^e_j = \mathrm{Embedding}^{\mathrm{e}}(s_j) \tag{1}$$

$$(\boldsymbol{e}_j, \boldsymbol{h}^e(j)) = \mathrm{GRU}^{\mathrm{e}}(\boldsymbol{x}^e_j, \boldsymbol{h}^e(j-1)) \tag{2}$$

$$\boldsymbol{h}^d(0) = \boldsymbol{h}^e(N_S) \tag{3}$$

The last hidden state of the encoder is passed to the decoder to initialize its hidden state as $\boldsymbol{h}^d(0)$. The first input to the decoder is a "<start>" index, and the output from the decoder is fed recurrently to the decoder as its next input. To output the kth token, a context vector $\boldsymbol{c}_k$ is obtained as a weighted sum of the encode vectors $\boldsymbol{e}_j$ with an attention weight vector $\boldsymbol{a}_k$. A GRU accepts a concatenation of an embedding of the last output token $\boldsymbol{x}^d_k$ and the context vector $\boldsymbol{c}_k$, and it updates the hidden state with passing a decode vector $\boldsymbol{d}_k$ to the fully-connected layer. The output token $\hat{t}_k$ is a word that has the maximum output probability at the fully-connected layer.

$$\boldsymbol{x}^d_k = \mathrm{Embedding}^{\mathrm{d}}(\hat{\boldsymbol{t}}_{k-1}) \tag{4}$$

$$\boldsymbol{a}_k = \mathrm{softmax}(\boldsymbol{v}^T \tanh(\boldsymbol{W}_1 \boldsymbol{E} + \boldsymbol{W}_2 \boldsymbol{H}^d_{k-1})) \tag{5}$$

$$\boldsymbol{c}_k = \sum_{j=1}^{N_s} a_k(j) \cdot \boldsymbol{e}_j \tag{6}$$

$$(\boldsymbol{d}_k, \boldsymbol{h}^d_k) = \mathrm{GRU}^{\mathrm{d}}(\boldsymbol{x}^d_k, \boldsymbol{c}_k) \tag{7}$$

$$\boldsymbol{y}_k = \mathrm{softmax}(\boldsymbol{W}_y \boldsymbol{d}_k + \boldsymbol{b}_y) \tag{8}$$

$$\hat{t}_k = \underset{v}{\mathrm{argmax}}(y_k(v)), \tag{9}$$

where E and H^d_{k-1} denote matrices of stacked encode vectors $\boldsymbol{e}_j (1 \le j \le N_S)$ and stacked duplications of a hidden state $\boldsymbol{h}^d_{k-1}$, respectively.

The model is trained in the normal sequence-to-sequence manner, where cross-entropy error is minimized by the stochastic gradient descent (SGD) algorithm.

$$loss = -\sum_{m=1}^{M} \sum_{k=1}^{\max N_T} \log y_k(t_k), \tag{10}$$

where M denotes the mini-batch size.

B. Decoding based on maximum mutual information criterion

As the base model tends to generate dull responses, we introduce decoding based on the MMI criterion. The decoder incorporates a penalty of Li's anti-LM with a decreasing weighting function [4]. The MMI-based decoder generates a response that maximizes the mutual information between S and T.

$$\begin{aligned}\hat{T} &= \underset{T}{\mathrm{argmax}} \left\{ \log \frac{p(S,T)}{p(S)p(T)} \right\} \\ &= \underset{T}{\mathrm{argmax}} \{\log p(T|S) - \log p(T)\}\end{aligned} \tag{11}$$

This equation is interpreted as penalizing frequent responses from the base of conditional log probabilities output by the base model. The MMI criterion is generalized with introduction of a hyperparameter λ for weighting the penalty. The penalizing term $\log p(T)$ is calculated as an accumulated log likelihood of a language model.

$$\hat{T} = \underset{T}{\mathrm{argmax}} \{\log p(T|S) - \lambda \log p(T)\} \tag{12}$$

$$\log p(T) = \sum_{k=1}^{N_T} \log p(t_k | t_{1:k-1}) \tag{13}$$

In practice, the 1-best sequence of output tokens is obtained by a greedy search. The $\boldsymbol{y}_k$ in (8) gives a set of conditional probability of the kth token t_k given the previous input S and a sequence of previous output tokens $\hat{t}_1, \ldots, \hat{t}_{k-1}$, and an n-gram language model gives the penalty. In Li's implementation, the penalty is only applied to tokens around the head of output responses because the penalty promotes ungrammatical sentences to be output as well. The MMI-based decoder determines an outputs token $\hat{t}_k$ as follows:

$$\hat{t}_k = \begin{cases} \underset{v}{\mathrm{argmax}}(\log y_k(v) - \lambda \log p(v|\hat{t}_{k-1})) & (k \le l) \\ \underset{v}{\mathrm{argmax}}(\log y_k(v)) & (k > l) \end{cases} \tag{14}$$

where l is the maximum length of penalizing frequent series of output tokens.

C. N-greedy search for diverse response generation

We extends the decoding with two types of N-best search. One is a basic beam search. The beam search keeps top K partial sentences with their accumulated scores every time-step for a search in the next time-step. From every partial sentence, conditional probabilities of the next tokens are computed, and the top L hypotheses are kept for comparison. Then, top K partial sentences are kept out of a set of $K \times L$ hypotheses. Finally, N-best sentences are selected on the basis of a mean score per token. A problem with the beam search is that the top K partial sentences are often occupied by those stretching from a common partial sentence when the number K is not very large.

The other N-best search keeps the top N words at the first (head) output token of the responses and greedily searches the

TABLE I
DATA SIZES OF FICTIVE CHITCHAT CORPUS.

Set	#pairs	#vocabulary	#tokens
Training	1,122,183	36,959	21,163,298
Validation	62,343	18,134	1,175,511
Test	62,344	18,104	1,177,674

1-best sequence of tokens following each of the N words at the head. To be exact, the first output token is selected as:

$$\{\hat{t}_{1,n}\}_{n=1,\ldots,N} = \text{N-}\max_{v}(\log y_1(v) - \lambda \log p(v|t_0)) \quad (15)$$

where t_0 is a "<start>" and the following tokens are determined by (14). The final N-best outputs are sorted on the basis of a mean score per token.

As the attention mechanism works best at the first output from the decoder, the N-best words at the head of sequences look substantially reasonable and diverse. Different words lead to diverse responses. We call this implementation N-greedy search.

IV. EXPERIMENTS

A. Experimental setup

The fictive chitchat corpus was separated into input-response pairs. The pairs were first divided into two classes corresponding to the personae. The Japanese sentences were preprocessed as follows:

1) Split every sentence into a sequence of words using the Japanese morphological analyzer "Mecab" [14].
2) Eliminate punctuation marks except for question marks and replace all words of only one occurrence in the corpus with "<oov>."
3) Insert a "<start>" and an "<end>" into the head and tail of every sentence, respectively, and convert every sentence into a sequence of indices in reference to a vocabulary table.

All pairs of one persona were split into training, validation, and test sets with a proportion of 18:1:1. The numbers of pairs in the training, validation, and test sets were 1.12M, 62k and 62k, respectively. The size of the vocabulary V, which was common to two personae, was 37,785. The rate of out-of-vocabulary words was 0.04%. The detailed numbers are listed in Table I.

A base model was trained for each of the personae using the normal sequence-to-sequence training algorithm that minimizes cross-entropy error with teacher forcing. The dimensions of the embeddings and hidden states were set at 128 and 256, respectively. The models were trained with 20 epochs. The language model used for calculating the penalty term was a unigram LM with additive smoothing. The language model was trained on the same corpus. The decreasing weighting function was set to penalize only the first word after "<start>", that is $l = 1$ in (14).

Three kinds of responses were generated for the common test set of inputs from the interlocutor using 1) the base model with a normal beam search (Base beam), 2) the MMI criterion with a normal beam search (MMI beam), and 3) the MMI criterion with the N-greedy search (MMI N-greedy). In the normal beam search, the numbers L and K were set at 5 to output 1-best and 3-best responses, while set at 10 to output more responses. In the N-greedy search, top N words were kept at the head to output N-best responses. The hyperparameter λ was tuned at 0.8, where Distinct-1 and Ent-1, two indices of diversity described below, reached their maximum levels with the validation set.

TABLE II
DIVERSITY INDICES FOR GENERATED RESPONSES.

	Dist-1	Dist-2	Ent-1	Ent-s	Len-s
1-best case					
Original	0.0249	0.1790	8.28	14.34	9.38
Base beam	0.0081	0.0232	5.57	7.40	7.17
MMI beam	0.0329	0.0534	7.52	10.88	4.95
MMI N-greedy	0.0274	0.0580	7.74	13.39	7.10
3-best case					
Original	0.0115	0.1010	8.27	14.67	9.37
Base beam	0.0017	0.0035	4.70	5.09	6.15
MMI beam	0.0131	0.0267	7.70	12.79	5.40
MMI N-greedy	0.0131	0.0342	7.78	14.59	7.16
5-best case					
Base beam	0.0014	0.0033	4.97	6.14	6.38
MMI beam	0.0082	0.0187	7.77	13.74	5.71
MMI N-greedy	0.0089	0.0262	7.81	15.12	7.18
10-best case					
Base beam	0.0009	0.0026	5.22	7.69	7.98
MMI beam	0.0040	0.0104	7.77	15.12	6.57
MMI N-greedy	0.0051	0.0178	7.84	15.85	7.24

The diversity of the responses was measured objectively with four indices: Distinct-1, Distinct-2, Ent-1, and Ent-s. Distinct-n is a ratio of the number of unique n-grams to the total number of n-gram tokens appearing in the corpus [4]. Dist-1 and Dist-2 stand for those of unigrams and bigrams in the test set. Ent-1 is the entropy of the unigrams in the test set with base 2 [7]. Likewise, Ent-s stands for sentence-level entropy. A low entropy means a biased distribution, that is a sign of frequent dull responses. Besides, the quality of responses was assessed subjectively from grammatical and contextual points of view.

B. Objective evaluation of diversity

Table II shows the four indices of diversity with an additional index Len-s for the test set in four cases: 1-, 3-, 5- and 10-best. The Len-s is the average number of tokens (length) in a response. The "Original" stands for the responses in the corpus. In the case of 1-best, one of three original responses to each input was randomly chosen.

Although all four indices of diversity decreased from the "Original" considerably with the "Base beam", the MMI-based response generation (MMI beam) significantly recovered all the indices. The N-greedy search (MMI N-greedy) further increased all the indices with Len-s increasing as well in all the case of the 3-, 5- and 10-best results. Table III lists an example of an input, three original responses, and generated 3-best responses with three methods.

TABLE III
EXAMPLES OF GENERATED RESPONSES (IN TRANSLATION).

Input		I like Hawaii. Europe is nice, too.
Original	1	Europe is nice. I want to visits quaint streets and musiums.
	2	In Europe, where do you want to visit?
	3	I like Hawaii, too. I want to surf a wave.
Base beam	1	Oh, I see.
	2	It sounds good. Great.
	3	It is very you.
MMI beam	1	Europe is nice. Want to visit there.
	2	Europe is nice. I want to visit there.
	3	Europe is nice. Want to go there.
MMI N-greedy	1	In Europe, where do you want to visit?
	2	Angkor Wat is also nice, isn't it?
	3	Hawaii is nice. I want to visit there, too.

C. Subjective assessment of quality

We conducted a subjective evaluation of the generated responses using a scale of discrete levels with crowd sourcing. A total of 6,000 input-response pairs were randomly sampled out of the test set except those including an "<oov>" in the input. Every input-response pair was scored by three raters on the scale of six levels listed in Table IV. The raters were collected through crowd sourcing. One rater scored 100 sets of three responses to a common input. The scores were tallied except 0 scores (responses to dull inputs), after checking if the rating was not conducted in an automatic manner.

Table V shows the percentages of the subjective scores and a ratio of the top among three methods including ties. The introduction of an MMI criterion greatly reduced the number of dull responses that occupied nearly half of all the responses to less than 10% and increased the ratio of the top from 58% to 60%. The N-greedy search achieved an additional reduction in the number of dull responses and an increase in the ratio of the top. However, a problem was that the ratios of score 1 and 2 increased more than those of score 4 and 5. This will be future work.

V. CONCLUSIONS

A non-task-oriented neural conversational model was trained with a large-scale fictive chitchat corpus in Japanese, and generated responses based on an MMI criterion was evaluated objectively in diversity and subjectively in quality. The four indices of diversity showed substantial increases by introducing the MMI criterion and additional increases by the efficient N-best search that keeps N-best words at the head of output sequences. The results of a subjective evaluation clearly showed a significant reduction in the number of dull responses and slight improvement in the general quality of the generated responses. The N-greedy search generated more diverse responses than the beam search while keeping the same level of quality.

While the number of dull responses was significantly reduced and the number of contextually correct responses increased, the number of contextually incorrect responses increased as well. This issue will be future work.

TABLE IV
CRITERION OF SUBJECTIVE RATING

Level	Criterion
5	Grammatically and contextually correct
4	Grammatically imperfect, but contextually correct
3	Dull response
2	Grammatically correct, but contextually incorrect
1	Grammatically and contextually incorrect
0	Dull input

TABLE V
PERCENTAGES OF SUBJECTIVE SCORES AND RATIO OF TOP AMONG THREE METHODS INCLUDING TIES.

	1	2	3	4	5	Top
Base beam	3%	10%	**47%**	7%	34%	58%
MMI beam	8%	27%	**9%**	9%	46%	60%
MMI N-greedy	8%	27%	**7%**	11%	47%	61%

REFERENCES

[1] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT 2015*, pp. 196–205, 2015.

[2] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of ACL-IJCNLP 2015*, pp. 1577–1586, 2015.

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

[4] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-prompting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016*, pp. 110–119, 2016.

[5] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *arXiv preprint, arXiv:1511.06732*, 2015.

[6] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. In *Proceedings of EMNLP 2016*, pp. 1192–1202, 2016.

[7] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of NIPS 2018*, pp. 1815–1825, 2018.

[8] Shohei Tanaka, Koichiro Yoshino, Sudoh Katsuhito, and Satoshi Nakamura. Conversational response re-ranking based on event causality and role factored tensor event embedding. In *arXiv preprint, arXiv:1906.09795*, 2019.

[9] Sam Wiseman and Alexander M Rush. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of EMNLP 2016*, pp. 1296–1306, 2016.

[10] Ilia Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of ICNLG 2019*, pp. 76–87, 2019.

[11] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *arXiv preprint, arXiv:1904.09751*, 2020.

[12] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP 2014*, pp. 1724–1734, 2014.

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

[14] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP 2004*, pp. 230–237, 2004.