フレーズ単位の発話応答ペアを用いた対話応答生成の多様化

佐藤 志貴 [1] 大内 啓樹 [2,1] 井之上 直也 [1,2] 鈴木 潤 [1,2] 乾 健太郎 [1,2]
[1] 東北大学 [2] 理化学研究所 AIP センター
{shiki.sato,naoya.inoue,jun.suzuki,inui}@ecei.tohoku.ac.jp
hiroki.ouchi@riken.jp

1 はじめに

対話応答生成は，ユーザの発話に対する適切な応答を生成することが目的のタスクである．近年，Sequence-to-sequence (seq2seq) モデル [1] を代表とするニューラル機械翻訳 (Neural Machine Translation; NMT) の枠組みを応答生成へ応用することによって，比較的流暢な応答を容易に生成可能となった．この応答生成法の問題点として，“I don’t know.” などの無難な単調応答 (dull responses) を頻繁に生成してしまうことが報告されている [2].

一方で，フレーズ (句) に基づく統計的機械翻訳 (Phrase-based Machine Translation; PBMT) [3] を応用した応答生成法も提案されている [4]．この手法では，学習データ内の入力発話に含まれるフレーズとその応答となるフレーズのペアを自動獲得し，それらを参照しながら適切な応答を生成する．実際の対話で用いられたフレーズペアを外部メモリのように保持・参照することにより，多様な応答が生成可能である．しかし，NMT に比べ，流暢さに欠けるという問題がある．

これら二つの手法の利点を活かすため，本研究では，NMT と PBMT を統合したハイブリッド生成モデル (図1) を対話応答生成に適用する．この手法では，まず PBMT 応答生成モデルが応答を生成する．次に，生成された応答と元の入力発話の両方を NMT モデルが入力として受け取り，応答を再生成する．これにより，NMT モデルによる応答生成の利点である発話の流暢さを保ちつつ，問題であった応答の多様化を図る．

評価実験として，NMT，PBMT，ハイブリッドモデルを応答の多様性の観点から評価した．全体としては NMT モデルに比べハイブリッドモデルの有意な改善は確認されなかったが，事例によってはハイブリッドモデルが PBMT モデルの出力を用いることで，NMT モデルよりも多様な応答生成をすることが確認された．

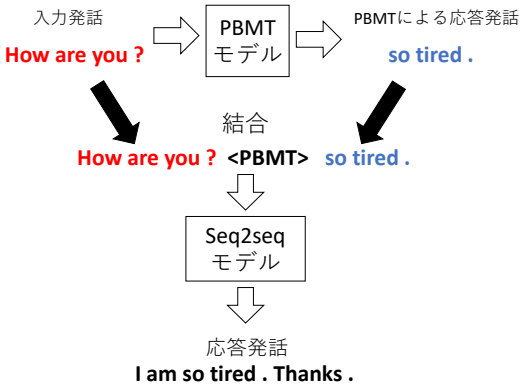

図1: PBMT-seq2seq ハイブリッドモデル

2 対話応答生成

本節では，対話応答タスクの設定とその評価指標について詳述する．

2.1 タスク設定

ユーザの発話 C に対して適切な応答 R を返すタスク (シングルターン対話応答生成) に取り組む．

$$\text{入力: } C = \{c_1, c_2, ...\}$$
$$\text{出力: } R = \{r_1, r_2, ...\}$$

例えば，ユーザの”How are you ?” という発話に対して”I am fine” という応答を生成した場合，$C = \{\text{How}, \text{are}, \text{you}, ?\}$，$R = \{\text{I}, \text{am}, \text{fine}\}$ となる．

2.2 評価指標

生成された応答発話 R に対して，Zhang らの先行研究 [5] を参考に以下 2 つの観点から評価する．

1. R が流暢かつ入力発話 C の応答として適切か
2. R が多様性に富んだ発話か

1 では R が文法上正しく，かつ C への応答として意味的に妥当かを評価する．例えば，”How are you?” という入力に対し”I like soccer.” という応答は文法上正しいが C への応答として意味的に妥当でないものとする．

生成される発話の意味的な妥当性を自動評価することは極めて難しい．そこで本研究では，データセットが提供する応答発話 (リファレンス発話) に近い応答発話は

流暢かつ入力発話への応答として適切なものとみなし，BLEU[6] を指標として用いる．また BLEU に加え，発話集合 $\mathcal{R}$ の中からリファレンス発話 R_{true} を選択する応答発話選択問題の正解率についても評価する．$\mathcal{R}$ は不正解としてテストデータ中からランダムサンプリングした 4 発話を加えた 5 発話の集合とする．発話選択は，モデルがタイムステップ t において応答発話中の単語 r_t をデコードする確率を $P(r_t)$ としたとき次式に従う:

$$\arg\max_{R\in\mathcal{R}} \frac{1}{m}\sum_{t=1}^{m} \log P(r_t|C, r_1, ..., r_{t-1}) \qquad (1)$$

2 では R が”I don’t know” などの，多様性を欠いた応答ではないかを評価する．多様性のない応答を生成しがちなシステムは，異なる入力発話に対しても似通った応答発話を生成しやすい．そのような応答を低く評価し，多様な応答を高く評価する自動評価指標がいくつか提案されている [2][7]．その中でも，本稿では最先端の自動評価指標 Ent-n[5] を用いる．Ent-n は，システムがどれだけ多様な n-gram を生成したかを，生成された各 n-gram の出現頻度情報を考慮して評価する指標である．

$$Ent = -\frac{1}{\sum_w F(w)}\sum_{w\in V} F(w)\log\frac{F(w)}{\sum_w F(w)} \qquad (2)$$

ここで，w は各 n-gram を表す．V はテスト中に出現した n-gram の集合であり，$F(w)$ は各 n-gram の評価データ中での頻度を表す．Ent-n は，高頻度 n-gram に対して低い値を割り当てるような指標となる．

3 従来法

3.1 NMT モデル

Sutskever らの seq2seq[1] はエンコーダー，デコーダーに Long short-term memory (LSTM) と呼ばれる再帰型ニューラルネットワーク (RNN) を用いた生成モデルである．エンコーダー側の LSTM で可変長の単語列を受け取り，デコーダー側の LSTM で可変長の単語列を順方向に 1 単語ずつ出力していく．

本研究では，エンコーダー側については入力単語列を順方向，逆方向の両方向から読むことで性能を向上させる双方向 LSTM[8] を用いる．またデコーダー側には，デコード時に入力単語列のどの部分を重視するかを考慮する Luong らのアテンション機構 [9] を用いる．

3.2 PBMT モデル

PBMT は訓練データから事前にフレーズ対とその翻訳確率を計算した上でフレーズテーブルを作成する．

表1: 抽出されたフレーズ対の例

	入力発話フレーズ	応答発話フレーズ
典型的	Thank you	welcome .
	do you like	I like
	I help you	I ’d like to
非典型的	he is pretty	and very considerate
	an aging populaton	increase the retirement
	How about winter ?	cold and damp

Ritter ら [4] によると PBMT モデルを直接対話応答生成システムとして用いる際に次のような問題が生じる．

1. 対話においては入力発話と出力発話の対の中で同一の句が繰り返されることが多い．そのため入力発話を繰り返すような応答を生成するモデルが学習される．
2. 入力発話と応答対話の対からは，機械翻訳モデルの学習に用いられるバイリンガルデータに比べてフレーズのアラインメントが取りにくい．

1 について同研究では，学習により抽出されたフレーズ対のうち一方のフレーズがもう一方のフレーズの部分単語列になるものをフレーズテーブルから除くことで対処した．またフレーズ対の各フレーズを単語集合としたときの Jaccard 係数を負の素性として追加した (Similarity 素性) ．

2 について同研究では，1 発話対のみからではなく訓練データ全体を参照してアラインメントをとることで対応した．具体的には，用意した大量のフレーズペアの候補に対し，訓練データ中での出現頻度をもとにフィッシャーの正確確率検定によってスコアリングし，上位のペアをフレーズテーブルに加えた．フレーズペアの候補は，全入力発話，応答発話の 1,2,3,4-gram 全てをそれぞれ入力発話フレーズ，応答発話フレーズとしたとき，一つ以上の訓練データ中の発話応答対で同時に出現するような組み合わせとする．表1にこの方法で得られたフレーズ対の例を示す．同表より”Thank you” - ”welcome .” のような対話のトピックに縛られず広く用いられるような出現頻度の高い典型的フレーズ対から，”How about winter ?” - ”cold and damp” などのトピックが限定された出現頻度の低い非典型的フレーズ対も得られた．

4 提案法

本研究では，PBMT モデルの生成する応答を NMT モデルが参照できるようにすることで，NMT モデルの流暢性と PBMT モデルの多様性を備えた応答生成手法を検討する．モデルは，機械翻訳においてニューラ

ルベースモデルが低頻度語の翻訳を正確に行うことなどを目的に Niehues らにより用いられた PBMT-NMT ハイブリッドモデル [10] を用いる．図1にモデルの概要を示す．3.1節の NMT モデルとの違いは，エンコーダーに対する入力として，3.2節の PBMT モデルが生成した C に対する応答発話 $R' = \{r'_1, ..., r'_{m'}\}$ を加えた点である．これにより NMT モデルが PBMT モデルの生成する応答発話を参照可能にした．エンコーダーには C と R' を特殊記号 <PBMT> でつないだ $\{c_1, ..., c_n, <\text{PBMT}>, r'_1, ..., r'_{m'}\}$ を入力する．

5 実験

4節で示した提案手法に対するベースラインモデルとして，(a) 3.1節で示した NMT 応答生成モデル (以下 `NMT-model` と表記)，(b) 3.2節で示した PBMT 応答生成モデル (以下 `PBMT-model` と表記) を用いた．これら `NMT-model`，`PBMT-model` および提案法を用いて 2章に示した実験を行い，その結果を比較・分析することで，提案法の特徴である NMT モデルが PBMT モデルの出力を参照可能とした場合の効果を検証した．

5.1 データセット

実験では雑談対話データセットである DailyDialog[11] を用いた．データセットに含まれる各対話データは複数発話から成るが，今回はシングルターンでの対話応答生成を行う．そのため訓練データおよび検証データについては，N 個の発話から成る 1 つの対話データから，隣接する発話同士を (入力発話，応答発話) として取り出して N-1 個のシングルターン対話データを作成した．テストデータにおいてはデータ中の各対話データの最終 2 発話を (入力発話，応答生成) とした．

シングルターン対話数は，訓練データ 76052，検証データ 7069，テストデータ 1000 対，1 発話あたりの単語数は 14.6 となった．

5.2 モデル設定

`PBMT-model` および提案手法の PBMT 部の実装には Moses[12] を用いた．Ritter ら [4] の設定に従いフレーズペア数は 5M，素性の重みは言語モデル 0.5，フレーズ翻訳モデル 0.2，Similarity を-0.2 とした．

`NMT-model` および提案手法の seq2seq 部の実装には Luong らのコードを用いた [13]．単語ベクトル，隠れ層を 300 次元として，単語ベクトルは GloVe[14] によって初期化し，語彙は GloVe の頻度上位 25000 語とした．エンコーダーは各方向 1 層の双方向 LSTM，デコーダーは 2 層 LSTM とした．また Dropout 確率は 0.2 とした．

表2: 自動評価結果

システム	BLEU	Ent-1	Ent-2	Ent-3	Ent-4	select
リファレンス	-	5.71	8.34	9.04	9.15	-
`NMT`	7.25	5.26	7.68	8.45	8.71	0.375
`PBMT`	3.20	5.38	8.00	8.92	9.20	-
提案手法	7.15	5.26	7.71	8.50	8.75	0.356

5.3 実験結果

表2に，各モデルによって生成された応答発話の評価結果を示す．なお，PBMT モデルでは各単語の生成確率が評価できないことから式 (1) が求まらないため，応答発話選択精度 (select) を評価しないこととした．

表2より，BLEU において `NMT-model` は 7.25 と `PBMT-model` のスコア 3.20 を大きく上回った．これにより NMT モデルが入力発話に対して妥当な応答を生成するタスクにおいて有効であることが改めて確認された．提案手法モデルは select において 0.356 と，`NMT-model` の 0.375 を下回った．しかし BLEU においては 7.15 と，`PBMT-model` の 3.20 と比べて `NMT-model` に近く，PBMT モデルの出力を用いることで NMT モデルの応答の適切さが大きく損なわれることはなかった．

一方で表2より，生成された発話の多様性においては `PBMT-model` がいずれの Ent-n においても `NMT-model` を上回っており，より多様な n-gram を生成していることが確認された．このことから，多様な応答発話の生成を目的としたとき，PBMT モデルの出力を用いることには一定の効果が見込まれることがわかった．しかし，提案モデルはどの Ent-n についても `NMT-model` のスコアをわずかに上回るのみで，`NMT-model` に比べて大きな多様性の向上は確認できなかった．

以上のように `NMT-model` と提案手法で BLEU および Ent-n のスコアに大きな差がないことから，提案手法が PBMT モデルの出力を参照せず，`NMT-model` に近い応答生成を行っている可能性がある．そのため各モデル出力の類似度を BLEU により評価した．表3に，各行に示したモデルの出力をリファレンスとしたときの，各列に示したモデルの出力の BLEU を示す．表3より，`NMT-model` の出力をリファレンスとした時の提案手法モデルの BLEU スコアはわずか 7.74 となり，両者の出力には一定の異なりがあることがわかった．この異なりについては，`PBMT-model` の出力をリファレンスとしたとき `NMT-model` の BLEU スコアは 1.97，比べて提案

表3: 各モデルの生成した応答の類似度 (BLEU)

	NMT	PBMT	提案手法
NMT	-	2.24	7.74
PBMT	1.97	-	5.45
提案手法	7.75	5.50	-

表4: 応答生成の出力例

入力発話	モデル	応答発話
Can I have another try?	提案手法	Sure. The fitting room is on the right.
	PBMT	Sure. The fitting room is that?
	NMT	Sure. Here you are.
Have they set a date for the wedding?	提案手法	No, not yet. But Susan says they'd like to have married in November or December.
	PBMT	in November or December.
	NMT	Not yet.

手法の BLEU スコアは 5.45 と高くなったため，提案手法モデルが PBMT モデルの出力の一部をモデルの出力に組み込むことを学習したことで生じた可能性がある．提案手法モデルが PBMT モデルの出力の一部を用いて応答生成をしたと思われる例を表4に示す．同表の例では，提案手法が PBMT 出力を用いることにより，NMT モデルに比べて多様な応答を生成したと考えられる．

6 関連研究

NMT ベースモデルによる多様な応答生成に向けて，先行研究では様々な方法が検討された [2][7]．中でも本研究に関連した研究として，Weston ら [15] は，検索モデルにより得られた応答を本研究における PBMT 出力と同じ要領で入力に加え，多様な応答生成を試みている．

フレーズペアを考慮した対話応答生成として，Wu らの研究が挙げられる [16]．同研究では，既存の対話ペアを用いて応答を生成するランキングモデルにおいて，フレーズペアを考慮したスコアリングを行っている．

7 おわりに

本研究では，対話応答生成タスクにおいて，PBMT と NMT を統合することによって，NMT 出力の特徴である流暢さを損なわない多様な応答生成ができるかを検証した．自動評価による多様性の向上は見られなかったが，事例によっては，提案手法が PBMT モデルの出力を用いることで NMT モデルの出力よりも多様な応答を生成していることがわかった．今後の課題として，PBMT モデルにより生成された応答発話の NMT モデルへの入力方法の改善や，NMT モデルへの入力として用いることを想定した PBMT モデルの工夫などが挙げられる．

謝辞

本研究の一部は JST 未来社会創造事業（JPMJMI17C7）の支援を受けて行った．

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. 2014, pp. 3104–3112.

[2] Jiwei Li et al. "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*. 2016, pp. 110–119.

[3] Philipp Koehn, Franz Josef Och, and Daniel Marcu. "Statistical Phrase-based Translation". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. NAACL '03. 2003, pp. 48–54.

[4] Alan Ritter, Colin Cherry, and William B. Dolan. "Data-Driven Response Generation in Social Media". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 2011, pp. 583–593.

[5] Yizhe Zhang et al. "Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. 2018, pp. 1815–1825.

[6] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[7] Kris Cao and Stephen Clark. "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*. 2017, pp. 182–187.

[8] M. Schuster and K. K. Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[9] Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 1412–1421.

[10] Jan Niehues et al. "Pre-Translation for Neural Machine Translation". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 1828–1836.

[11] Yanran Li et al. "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2017, pp. 986–995.

[12] Philipp Koehn et al. "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. 2007, pp. 177–180.

[13] Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. "Neural Machine Translation (seq2seq) Tutorial". In: *https://github.com/tensorflow/nmt* (2017).

[14] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.

[15] Jason Weston, Emily Dinan, and Alexander Miller. "Retrieve and Refine: Improved Sequence Generation Models For Dialogue". In: *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. 2018, pp. 87–92.

[16] Xianchao Wu et al. "りんな：女子高生人工知能". In: 言語処理学会第 *22* 回年次大会. 2016.