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
東北大学 工学部 電気情報物理工学科

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流暢性・多様性の観点から評価した．ハイブリッドモデルが，NMT と同程度の流暢さを保ちながらも，PBMT に比べ NMT の多様性が低くなるような状況において NMT より多様な応答を生成することが確認された．

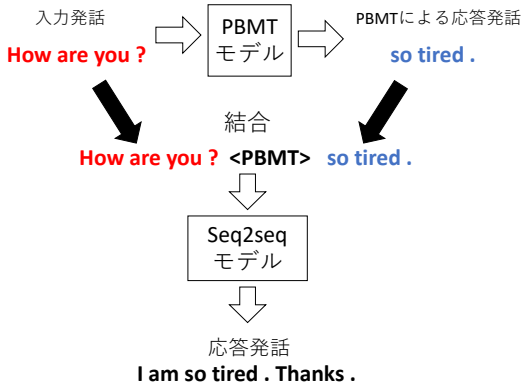

図1: PBMT-seq2seq ハイブリッドモデル

2 対話応答生成

本章では，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．しかし，対話応答生成に

おいてはリファレンス発話とモデル出力での 3-gram や 4-gram の一致率が極めて低く，評価尺度として用いることは難しい．そのため，1-gram，2-gram の一致率を評価する BLEU-1，BLEU-2 のみを用いる．

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，本研究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．

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

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．BLEU と同様に，1-gram と 2-gram の多様性を評価する Ent-1，Ent-2 を用いる．

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．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 DailyDialog[10] の訓練データから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．

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

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 [11] 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

実験には，DailyDialog および Cornell Movie-Dialogs Corpus[12] を用いた．

5.1.1 DailyDialog

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 76,052 対，検証データ 7,069 対，テストデータ 1,000 対となった．

5.1.2 Cornell Movie-Dialogs Corpus

本データセットは 617 タイトルの映画スクリプトから 2 話者間の対話データを抜き出したデータセットとなっている．映画タイトルをランダムに 495，61，61 に分割し，それぞれを訓練データ，検証データ，テストデータとした．実験には，DailyDialog の訓練データ，検証データと同じ方法で各対話データをシングルターン対話データに分解して用いた．データ中には極端に長い発話が含まれていたため，入力発話と応答発話の単語数がどちらも 20 以下であるようなデータのみを用いた．また，データ中にリメイク作品のスクリプトなどが含まれていた影響で重複した対話データが含まれていたため，これらの除去を行った．

シングルターン対話数は，訓練データ 115,509 対，検証データ 14,907 対，テストデータ 15,270 対となった．

表2: 自動評価結果 (BLEU)

	DailyDialog		Cornell	
モデル	BLEU-1	BLEU-2	BLEU-1	BLEU-2
`NMT`	26.7	9.45	18.7	1.86
`PBMT`	17.8	3.57	14.8	1.00
提案手法	28.3	10.2	19.4	1.78

表3: 自動評価結果 (Ent)

	DailyDialog		Cornell	
モデル	Ent-1	Ent-2	Ent-1	Ent-2
リファレンス	5.71	8.34	5.71	9.29
`NMT`	5.26	7.68	2.79	3.20
`PBMT`	5.38	8.00	5.21	8.50
提案手法	5.26	7.71	4.13	6.16

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[13] 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 [14]．単語ベクトル，隠れ層を 300 次元として，単語ベクトルは GloVe[15] によって初期化した．DailyDialog を用いた実験においては GloVe の頻度上位 25,000 語を語彙とした．また Cornell Movie-Dialogs Corpus を用いた実験においては，学習データ中に 3 回以上出現し，かつ GloVe により初期化可能な 17,194 語を語彙とした．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．

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 BLEU-1，BLEU-2 のスコアを示す．また，表3に Ent-1，Ent-2 のスコアを示す．

表2より，データセットとして DailyDialog を用いた場合と Cornell Movie-Dialogs Corpus (Cornell) を用いた場合，ともに BLEU において `NMT-model` が `PBMT-model` を上回った．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．また，提案手法の BLEU スコアは両データセットを用いた場合ともに

NMT-model と同程度となった．このことから，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損なわれることはないということがわかった．

一方で表3より，生成された発話の多様性においては PBMT-model がどちらのデータセットを用いた場合においても NMT-model を上回った．ただし，Cornell Movie-Dialogs Corpus を用いた際には PBMT-model のスコアが NMT-model を大きく上回った一方で，DailyDialog を用いた際にはその差は僅かなものとなった．提案手法の Ent-n スコアにおいては，DailyDialog を用いた場合 NMT-model に比べ有意な改善が見られなかった．一方で，Cornell Movie-Dialogs Corpus を用いた場合，提案手法の Ent-n スコアが NMT-model を大きく上回ることがわかった．

5.4 出力の分析

実験において各手法により生成された応答を比較するために，BLEU スコアを用いて各手法の出力の類似度を測定した．表4に，DailyDialog を用いて実験を行った際の出力について，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 スコアを示す．また表5に，Cornell Movie-Dialogs Corpus を用いて実験を行った際の出力について，表4と同様の測定結果を示す．表4および表5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 スコアはどちらのデータセットについても 10 程度にとどまり，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 スコアに比べて提案手法の BLEU スコアが高くなったため，提案手法が PBMT モデルの出力の一部を応答に組み込んだことで生じた可能性がある．このことから5.3節において実験に DailyDialog を用いた場合のように，PBMT と NMT による出力の多様性に大きな差が見られないとき，PBMT モデルの応答を組み込む提案手法では多様性が改善されないと考えられる．一方で，実験に Cornell Movie-Dialogs Corpus を用いた場合のように NMT が PBMT の多様性を大きく下回るときには，提案手法が PBMT の出力を組み込んだ応答を生成することで，NMT に比べ多様な応答を生成したと思われる．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6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．

表4: 各モデルの生成した応答の BLEU (DailyDialog)

	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5: 各モデルの生成した応答の BLEU (Cornell)

	NMT	PBMT	提案手法
NMT	-	2.91	10.4
PBMT	3.00	-	18.9
提案手法	10.0	18.3	-

表6: 応答生成の出力例

入力発話	モデル	応答発話
Open the door, will ya? (Cornell)	提案手法	I can't open it.
	PBMT	I can't open be '.
	NMT	I don't know.
But it looks good. Let me try it on. (DailyDialog)	提案手法	that's the latest model
	PBMT	Yes, it's the latest model - over there?
	NMT	you 're welcome

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．Li らによる研究 [2] や中村らによる研究 [16] では，モデル学習時の目的関数を改善することで単調な応答の生成を抑制するという手法を用いた．また，Cao ら [7] の，潜在変数を導入することによる応答生成の多様化なども挙げられる．本研究はこれらと併用可能なアプローチとなっている．

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 [17] が挙げられる．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り，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．提案手法が PBMT の応答を組み込むような形で応答を生成することで，NMT と同程度の多様性を有しながらも，NMT の多様性が PBMT の多様性を下回る場合において NMT よりも多様な応答を生成することが確認された．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] 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.

[11] 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.

[12] Cristian Danescu-Niculescu-Mizil and Lillian Lee. “Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs”. In: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, 2011, pp. 76–87.

[13] 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.

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

[15] 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.

[16] Ryo Nakamura et al. “Another Diversity-Promoting Objective Function for Neural Dialogue Generation”. In: *AAAI 2019 Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL 2019)*. 2019.

[17] 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.