句の呼応と話題の一貫性に着目した低品質対話データの教師なしフィルタリング

赤間 怜奈 [1,2]　鈴木 潤 [1,2]　横井 祥 [1,2]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
{reina.a, jun.suzuki, yokoi, inui}@ecei.tohoku.ac.jp

1 はじめに

深層ニューラルネットワーク（DNN）技術は，自然言語処理における文生成の発展を大きく牽引してきた．DNN を用いた機械翻訳分野では，モデルの改善と同様に，あるいはそれ以上に大規模で**高品質な訓練データ**が性能向上に大きく寄与することが実証されている [8]．

対話応答生成分野でも SNS や映画字幕などの大規模対話データがニューラル対話応答生成の発展に貢献してきたが，これらは一般に高品質とは言い難い．たとえば，研究用途で最も広く使用される大規模対話コーパス OpenSubtitles [10] は，慣習的に，話者情報が付与されていない映画字幕の 1 行分が 1 発話に相当するという粗い仮定の下で，連続する 2 発話を発話–応答ペアと見做し対話データを獲得する．当然ながら，これには対話として明らかに許容できない低品質な発話–応答ペア[*1]が含まれること（表1参照）が指摘されている [3, 9]．しかし，こうした低品質な発話-応答ペアを効果的に除去する方法論は確立されておらず，従来研究ではこれらを含んだままのデータを用いて研究がおこなわれてきた．

本研究では，低品質な発話–応答ペアを除去することを目的として，各発話–応答ペアに対し対話としての品質の良さを算出するスコア関数を提案する．提案する関数は，対話研究で「良い対話」の要素として言及されてきたふたつの観点 (i) 発話と応答が対話らしく自然に繋がっているか，(ii) 発話と応答の内容が互いに関連しているか，をスコアとして反映するように設計する．提案法は，特定の教師データを必要とせず，またドメインや言語などデータの特性に依存しないため，任意の対話データに適用可能である．

実験では，提案法により算出したスコアが人間の主観と相関を持つことを示す．また，提案法を用いて低品質な発話–応答ペアを除去する訓練データフィルタリングが対話応答生成モデルの性能向上に有効であることを，自動評価および人手評価により検証する．

[*1] シーンを跨ぐ無関係の発話同士がペアとして繋がってしまう，等．

2 発話–応答ペアの良さをどう判断するか？

本研究の目的は，対話データから低品質な発話–応答ペアを自動的に検知し，除去することである．発話–応答ペアの対話としての良さの計算方法を，既存研究における人手評価基準と，実データの観察を通して探る．

2.1 既存研究における人手評価基準

連続するふたつの発話が与えられたとき，人間は何をもって「良い対話である」と判断するのだろうか．我々は，対話応答生成タスクにおける人手評価の観点を参考に，その判断基準を探る．調査の結果，対話の良さに寄与する観点は以下のふたつに集約さることがわかった．

ひとつめは，先行する発話に対して応答が対話らしく自然に（適切に）繋がっていることである（以下，**対話らしい繋がり**と表記）．評価者に発話–応答ペアの対話としての良さを判断させる際，Shang らは *“an appropriate and natural response to the post”* [16] を，また Xing らは *“the response can be used as a reply”* [20] を基準とするよう指示した．他にも様々な先行研究が同様の観点，*“natural to”*, *“appropriate for”* や *“coherent with”* から対話の良さを判断するよう求めていた [1, 11, 12, 17]．

ふたつめは，発話と応答の内容が互いに関連していることである（以下，**内容の関連性**と表記）．Galley らは *“evaluate responses in terms of their relevance”* を評価者に求めた [7]．発話と応答間の *relevance* を評価する研究は多い [11, 12, 21]．また，Li らは先行する発話に対し *“more specific to”* な応答を高く評価するよう評価者に指示した [9]．Ritter らは *“an appropriate response should be on the same topic”* と評価者に教示した [13]．

自然言語処理分野で対話応答生成の人手評価基準を集約したこのふたつの観点は，社会言語学分野でも対話の重要な要素と考えられている [14, 18]．

2.2 実際の対話についての観察

前節で述べたふたつ観点が対話データにどのような形で現れるかを，人手評価[*2]を通して観察する（表1）．

[*2] OpenSubtitles から作成した発話–応答ペアのうち無作為に抽出した 100 ペアについて，4.1節と同様の方法で人手評価を実施した．評価

表1　人手評価で高評価または低評価を得た発話–応答ペアの例．青色部は対話らしい繋がりに寄与すると考えられる発話応答間で呼応関係にある句を表す．各発話および応答から推定されるトピックを文末に記す．

	発話	応答	人手評価
1：	It'll be like *you* never left. [topic:??]	*I* painted a white line on the street way over there. [topic:painting]	1.4
2：	You're gonna get us assimilated. [topic:??]	Switch to a garlic shampoo. [topic:??]	1.8
3：	*I* probably *asked for* too much money. [topic:money]	Money's always a problem, *isn't it?* [topic:money]	4.2
4：	*You've been* borderline stalking Angela as long as *we've been friends*. [topic:friendship]	*We've been friends* since we were five. [topic:friendship]	4.6

まず**対話らしい繋がり**について，人手評価で高評価を得たペアでは，発話内の特定の句に呼応する句が応答内に登場した．たとえば，表1の 3 行目では，何かを依頼する表現 "*ask for*" とそれに対し確認を返す表現 "*isn't it?*" が呼応関係にある．他にも (*why*, *because*) や (*what do you want*, *I want*) など，たとえば言語学分野の cohesive devices [15] のような，対話における典型的な応酬が句の対応として現れる傾向にあった．

次に**内容の関連性**について，高評価を得たペアでは，多くの場合で発話と応答の両方が同一のトピックについて言及していた．共通の場面や話題を示唆する表現は，文全体に渡って観察された．

以上より，本論文では (i) 発話応答間で呼応する句の存在，(ii) 発話と応答それぞれが言及するトピックの共通性を基準に，対話の品質が計算可能という仮説を置く．

3　提案: スコア関数の設計

前章の仮説に基づき，対話らしい繋がりと内容の関連性の両方を尺度として反映するスコア関数を提案する．

問題設定　n 組の発話–応答ペア集合を対話データ $\mathcal{D}$ とおく．$\mathcal{D}$ は低品質な発話–応答ペアも含む．

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}. \tag{1}$$

我々は，各発話–応答ペア $(x_i, y_i) \in \mathcal{D}$ に対し，品質の良さをスコアとして与える関数 $S_{\mathbf{I}+\mathbf{R}}$ を設計する．

対話らしい繋がり　呼応する典型的な句ペアの存在を手がかりに，連続する 2 発話 (x, y) が対話らしく繋がっているかを，$S_{\mathbf{I}}(x, y)$ として，計算する．

まずはじめに，$\mathcal{D}$ に含まれる「呼応する典型的な句ペアの集合」$\mathcal{P} = \{(f_j, e_j)\}$ を獲得する[*3]．$\mathcal{P}$ の獲得には，統計的機械翻訳研究の IBM モデルに基づくフレーズテーブル抽出技術，たとえば Moses[*4]を利用する．統計的機械翻訳ではできるだけ全ての句の間に一対一の対応をつけることが目的となるが，前節の観察より，対話の繋がりに寄与するのは一部の句ペアに過ぎない．この性質を考慮し，実験では，IBM モデル上の null-alignment 確率を 0.5 に設定し，またフレーズ抽出アルゴリズム上の句の範囲を貪欲に広げる設定を解除した上で，フレーズテーブルを作成した．

次に，呼応する句ペアの集合 $\mathcal{P}$ を用いて，発話–応答ペア (x, y) 繋がりの自然さ（どの程度多くの呼応する句ペアを含むか）を算出する:

$$S_{\mathbf{I}}(x, y) := \sum_{(f,e)\in\phi(x,y)\cap\mathcal{P}} \mathrm{nPMI}(f, e) \times \frac{|f|}{|x|} \times \frac{|e|}{|y|}. \tag{2}$$

ここで，$\phi(x, y)$ は文ペア (x, y) から抽出可能なすべての句（n-gram）ペアの集合，$|\cdot|$ は発話または句に含まれるトークン数を表す．正規化された自己相互情報量（nPMI） [4] は，低頻度の句ペアが極端に大きな情報量を持つことを防ぐ．また，$|f|/|x|$，$|e|/|y|$ によって，呼応する句ペアが文に占める割合をスコアに反映する．

内容の関連性　トピックの共通性を手がかりに，ふたつの発話 x, y の内容の関連性 $S_{\mathbf{R}}(x, y)$ を計算する:

$$S_{\mathbf{R}}(x, y) := \cos(\boldsymbol{v}(x), \boldsymbol{v}(y)). \tag{3}$$

ここで，$\boldsymbol{v}(\cdot)$ は発話に対応するベクトル表現（文ベクトル）を表す．2 文間の関連性は，文を構成する単語の単語ベクトルから文ベクトルを構築し，それらのコサイン類似度で計算できることが既存研究で示されている [5, 19]．対話応答生成においても発話応答間の関連性を考慮する手段としてしばしば用いられる [21]．

まとめ　ふたつの観点の足し合わせにより，最終的なスコアを算出する:

$$S_{\mathbf{I}+\mathbf{R}}(x, y) := \alpha S_{\mathbf{I}}(x, y) + \beta S_{\mathbf{R}}(x, y). \tag{4}$$

ハイパーパラメータ $\alpha, \beta \in \mathbb{R}_{\geq 0}$ は，各観点の重みを決定する．我々の実験では，ふたつの観点が同じスケールの下で足し合わさるよう，次のように定めた:

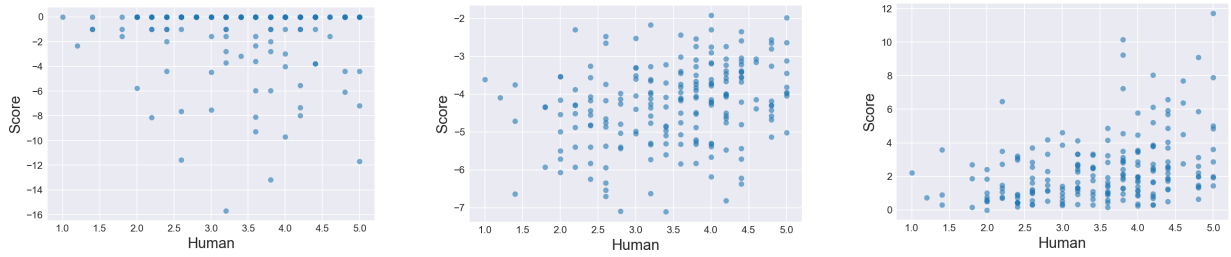
$$\alpha = 1/\left(\tfrac{1}{n}\textstyle\sum_{(x,y)\in\mathcal{D}} S_{\mathbf{I}}(x, y)\right), \tag{5}$$

$$\beta = 1/\left(\tfrac{1}{n}\textstyle\sum_{(x,y)\in\mathcal{D}} S_{\mathbf{R}}(x, y)\right). \tag{6}$$

者は「対話として許容できるか」を直感的に判断した．

[*3] 各フレーズペア (f_j, e_j) は，いずれかの文ペア $(x_i, y_i) \in \mathcal{D}$ に含まれるとする：$f_j \in x_i$，$e_j \in y_i$．

[*4] http://www.statmt.org/moses/.



(a) Entropy TRG [6]　(b) Dual-cond. [8]　(c) 提案 $S_{\mathbf{I+R}}$

図1　各手法で算出したスコア（縦軸）と人手評価スコア（横軸）の関係．左下が低評価，右上が高評価を表す．

4　実験

提案したスコア関数が (1) 人間の主観と相関を持つこと (4.1節)，(2) 対話応答生成のための訓練データフィルタリングに有効であること (4.2節) をそれぞれ示す．

4.1　対話データのスコアリング

発話–応答ペアの品質評価において，提案スコアが人間の主観と相関を持つことを示す．

データセット　本研究では，対話データ $\mathcal{D}$ として OpenSubtitles [10]（英語字幕約 441M 行）から作成した約 80M の発話–応答ペア集合を用いた[*5]．

スコア算出　繋がりの自然さ $S_{\mathbf{I}}$ の算出のために，統計的機械翻訳ツール Moses と $\mathcal{D}$ を用いてフレーズテーブルを学習し，句ペア集合 $\mathcal{P}$ を獲得した．このとき，アライメントポイントの計算には FastAlign[*6]を採用した．共起頻度が 200 未満の句ペア (e, f) はテーブルから除去し，最終的な句ペア集合 $|\mathcal{P}| = 68{,}891$ を獲得した．内容の関連性 $S_{\mathbf{R}}$ を算出する際には，単語ベクトルとして訓練済み fastText[*7]を用い，Arora らの手法 [2] により文ベクトルを作成した．彼らの手法は 2 文の関連度合いを計算するタスクにおいて高い精度を達成しており [19]，本研究の 2 発話間のトピックの共通性を判断するという目的に適する．$\mathcal{D}$ 内の全ての発話–応答ペアについて，提案スコア $S_{\mathbf{I+R}}$ を算出した（表3）．

比較手法　提案法の効果を相対的に検証する目的で，実験では以下 2 通りの既存手法を用いた結果も報告する:

- Entropy: 発話の汎用性をエントロピーとして算出し，この値が大きいものを訓練データから除くことで多様性のある対話応答生成を実現 [6].
- Dual-cond.: エンコーダデコーダモデルを用いて双方向条件付きクロスエントロピーに基づくスコアを算出し，これが小さいものを訓練データから除くことで翻訳性能の向上を実現 [8]．WMT2018 の Noisy Parallel Corpus Filtering Task で最高性能を達成．

表2　各手法が算出したスコアと人手評価スコアの相関．

スコアリング手法	Spearman's ρ	p 値
Entropy SRC [6]	-0.1173	9.8×10^{-2}
Entropy TRG [6]	0.0462	5.2×10^{-1}
Dual-cond. [8]	0.2973	1.9×10^{-5}
提案 $S_{\mathbf{I+R}}$	**0.3751**	$\mathbf{4.4 \times 10^{-8}}$
提案 $S_{\mathbf{I}}$	0.2044	3.7×10^{-3}
提案 $S_{\mathbf{R}}$	0.3007	1.5×10^{-5}

評価結果　提案法および比較手法が発話–応答ペアについて算出したスコアと人手評価スコアとの関係を図1に，相関係数を表2に，それぞれ示す．評価には $\mathcal{D}$ から無作為に抽出した 200 ペアを用いた．各ペアについて 5 人の評価者が「連続するふたつの発話が対話として許容できるか」の問いに 5 段階で回答し[*8]，その平均値を人手評価スコアとした．表2より，提案法はペアの品質評価において人間の主観と最も高い相関があった．また，図1(c) より，提案法はペアの品質を人間の主観より低く見積もることはあれど高く見積もることは少ない，という傾向にある．これは，低品質なペアを確実に除去することが求められるフィルタリングに適した性質と考えられる．

4.2　対話応答生成のための訓練データフィルタリング

事例研究として，提案法が訓練データフィルタリングに有効であることを示す．

低品質ペアの除去　提案法により検知した訓練データ中の低品質な発話–応答ペアを除去することで，訓練データの品質向上と，これに起因する生成モデルのパフォーマンスの向上が期待できる．対話データ $\mathcal{D}$ のうち提案法によるスコアリングで下位約 50% に該当した発話–応答ペアを除去し，40,000,000 ペアから成るフィルタリング済み訓練データを獲得した．さらに，前述の比較手法たち（4.1節参照）を用いて $\mathcal{D}$ から獲得したフィルタリング済み訓練データも，それぞれ同規模で準備した．

[*5]前処理として，他言語の発話・オウム返し・ペアの重複を除去した．
[*6]https://github.com/clab/fast_align.
[*7]https://fasttext.cc/.
[*8]スコア 1: strongly disagree〜5: strongly agree の 5 段階．評価には Amazon Mechanical Turk を利用．評価者は英語母語話者に限定．

表3　発話–応答ペアについて提案法が算出したスコアと人手評価スコアの例．$S_{\mathbf{I}}$ および $S_{\mathbf{R}}$ は α，β により正規化済み．

	発話	応答	$S_{\mathbf{I}}$	$S_{\mathbf{R}}$	$S_{\mathbf{I+R}}$	人手評価
1 :	Pushers won't let the junkie go free.	Across 110th Street.	0.00	0.42	0.42	2.4
2 :	It started when I was 17.	They'd make a cash drop,	0.63	0.00	0.63	2.0
3 :	A big nail should be put in your head	Who are they	0.74	0.00	0.74	1.2
4 :	He told me so.	Oh, he did, huh?	2.21	0.00	2.21	4.8
5 :	Then if I win, what are you going to do?	When you win?	1.04	7.01	8.05	4.2
6 :	But what do you want me to do?	We want you to kick her off the team.	10.20	1.53	11.72	5.0

表4　生成応答の自動評価および人手評価結果．**太字**は各評価尺度における最良の結果を表す．

フィルタリング手法	訓練データサイズ	平均単語長 (差分)	distinct-1	distinct-2	BLEU-1	ROUGE	人手評価
（フィルタリングなし）	79,445,453	8.44 (−0.60)	127/0.030	238/0.064	8.8	7.71	3.37
無作為サンプル	40,000,000	**8.66 (−0.38)**	123/0.028	223/0.058	9.1	8.10	3.36
Entropy SRC [6]	40,000,000	7.97 (−1.07)	165/0.041	329/0.094	9.1	7.76	3.56
Entropy TRG [6]	40,000,000	18.25 (+9.21)	213/0.023	591/0.069	5.4	6.86	2.85
Dual-cond. [8]	40,000,000	8.63 (−0.41)	206/0.048	478/0.125	**9.4**	**8.32**	3.43
提案 $S_{\mathbf{I+R}}$	40,000,000	7.13 (−1.91)	**345/0.097**	**853/0.278**	**9.4**	7.50	**3.73**
参照応答		9.04 (±0.00)	1301/0.288	3244/0.807	-	-	-

学習設定　応答生成モデルには Transformer を使用した[*9]．トークン分割には Byte Pair Encoding を用いた[*10]．

評価結果　それぞれの訓練データで学習したモデルにより生成された応答について，自動評価および人手評価の結果を表4に示す．自動評価および人手評価には，テストセットから無作為に抽出したそれぞれ 500 および 100 の発話–応答ペアを用いた．提案法によるフィルタリング済み訓練データで学習したモデルは，生成応答の異なり n-gram 数/率を表す自動評価尺度 distinct-n および人手評価において，フィルタリングなしの場合よりも良い結果となり，さらには他のフィルタリング手法と比較しても最良の結果となった．以上より，提案法による訓練データフィルタリングは，多様性を保持した適切な応答生成の促進に有効であることが実験的に示された．

5　おわりに

本研究では，ニューラル対話応答生成のの訓練データに含まれる低品質な発話–応答ペアを検知するためのスコア関数を提案した．提案法は，発話–応答ペアに対し「対話らしく自然に繋がるか」と「内容的に関連するか」のふたつの観点を反映したスコアを与える．提案法はドメインや言語などデータの特性に依存しないため，任意の対話データに適用できる．実験では，発話–応答ペアの品質評価において提案法のスコアが人間の主観と相関を持つこと，提案スコアによる訓練データフィルタリングはニューラル対話応答生成モデルの性能向上に有効であることをそれぞれ示した．今後は異なるドメインや言語の対話コーパスにおける提案法の効果を検証したい．

謝辞　本研究は JSPS 科研費 JP19H04162，JP19J21913 の助成を受けたものです．

[*9] https://github.com/pytorch/fairseq.
学習パラメータは'`--arch transformer_wmt_en_de_big`' のデフォルト設定．最大訓練ステップ数 100K．生成時のビーム幅 10.

[*10] https://github.com/rsennrich/subword-nmt．語彙数 16K.

参考文献

[1] R. Akama et al. "Generating Stylistically Consistent Dialog Responses with Transfer Learning". In: *IJCNLP*. Vol. 2. 2017, pp. 408–412.
[2] S. Arora et al. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings". In: *ICLR*. 2017.
[3] A. Baheti et al. "Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints". In: *EMNLP*. 2018, pp. 3970–3980.
[4] G. Bouma. "Normalized (Pointwise) Mutual Information in Collocation Extraction". In: *GSCL*. 2009, pp. 31–40.
[5] A. Conneau et al. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *EMNLP*. 2017, pp. 670–680.
[6] R. Csáky et al. "Improving Neural Conversational Models with Entropy-Based Data Filtering". In: *ACL*. 2019, pp. 5650–5669.
[7] M. Galley et al. "deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets". In: *ACL*. Vol. 2. 2015, pp. 445–450.
[8] M. Junczys-Dowmunt. "Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora". In: *WMT*. 2018, pp. 888–895.
[9] J. Li et al. "A Diversity-Promoting Objective Function for Neural Conversation Models". In: *NAACL*. 2016, pp. 110–119.
[10] P. Lison et al. "OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora". In: *LREC*. 2018, pp. 1742–1748.
[11] R. Lowe et al. "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses". In: *ACL*. Vol. 1. 2017, pp. 1116–1126.
[12] J. Pei et al. "S2SPMN: A Simple and Effective Framework for Response Generation with Relevant Information". In: *EMNLP*. 2018, pp. 745–750.
[13] A. Ritter et al. "Data-Driven Response Generation in Social Media". In: *EMNLP*. 2011, pp. 583–593.
[14] H. Sacks. "Lecture One: Rules of Conversational Sequence". In: *Human Studies* 12.3/4 (1989), pp. 217–233.
[15] R. Salkie. *Text and Discourse Analysis*. Routledge, 1995.
[16] L. Shang et al. "Neural Responding Machine for Short-Text Conversation". In: *ACL*. Vol. 1. 2015, pp. 1577–1586.
[17] X. Shen et al. "A Conditional Variational Framework for Dialog Generation". In: *ACL*. Vol. 2. 2017, pp. 504–509.
[18] J. Sidnell. *Conversation Analysis: An Introduction*. Language in Society. John Wiley & Sons, 2010.
[19] S. Subramanian et al. "Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning". In: *ICLR*. 2018.
[20] C. Xing et al. "Topic Aware Neural Response Generation". In: *AAAI*. 2017, pp. 3351–3357.
[21] X. Xu et al. "Better Conversations by Modeling, Filtering, and Optimizing for Coherence and Diversity". In: *EMNLP*. 2018, pp. 3981–3991.