エントレインメントスコアを用いた応答リランキングとその自動評価

金崎 翔大 [1,2]　河野 誠也 [2]　湯口 彰重 [2]　桂井 麻里衣 [3]　吉野 幸一郎 [2]
[1] 同志社大学大学院理工学研究科　[2] 理化学研究所ガーディアンロボットプロジェクト
[3] 同志社大学理工学部
{kanezaki21,katsurai}@mm.doshisha.ac.jp
{seiya.kawano,akishige.yuguchi,koichiro.yoshino}@riken.jp

概要

人間と人間の会話では，エントレインメントと呼ばれる対話の進行に従って話し方が同調する現象がしばしば発生する．著者らは直近の研究で，対話文脈に対して理想的なエントレインメント度合いを自動決定し，それに基づきニューラル雑談対話モデルの応答リランキングをする手法を提案した．本研究ではこの枠組みに基づき，新たに BERTScore に基づくエントレイメント尺度を提案する．また，従来研究で提案したエントレインメント尺度と本研究で提案したエントレイメント尺度に基づいて，ニューラル雑談対話モデルの応答リランキングを行う．また，このリランキング結果に対する自動評価を行う．

1　はじめに

エントレインメントは，収束やアラインメントとも呼ばれ，対話における話者間の話し方が対話の進行に従い類似する現象を指す．この現象は，語彙 [1]，統語構造 [2]，文体 [3]，韻律 [4]，ターンテイキング [5]，対話行為 [6] など，対話における様々な要素で観測される．エントレインメントは，対話のタスク成功率や自然性，対話意欲と相関することが報告されており，エントレインメントの分析 [7, 8] を通して対話システムの性能や対話の質を評価する試みが行われている．また，エントレインメントを考慮して，対話システムの応答選択・応答生成を制御する試みもある [9, 10, 11]．

著者らは直近の研究で，対話文脈に対する応答のエントレインメント予測モデルを構築し [12]，それに基づきニューラル雑談対話モデルの n-best 応答候補をリランキングする手法を提案した [11]．この研究では，システムが対話文脈に対して “どの程度” エントレインメントした応答をするべきかを予測モデルによって学習し，“理想的” なエントレインメント度合いをもたらすような応答を自動決定する手法を提案した．提案したリランキング手法が主観評価を顕著に悪化させることなく応答のエントレインメントを制御することが可能であることを示した．一方で，この研究で用いた単語分散表現に基づく尺度だけではエントレインメント現象を十分に捉えられず，対話システムの応答生成に利用するための十分な情報量を持っていない可能性が示唆された．

言語的なエントレインメント分析における先行研究では，様々なエントレインメント尺度が提案されている．Nenkova らは対話コーパス内の頻出単語に着目し，2 名の話者によるそれらの利用傾向に基づくエントレインメント尺度を提案した [7]．しかし，頻出単語のみに着目した尺度では，単語の意味や統語構造などの複数の側面を捉えきれない可能性があった．そこで，Nasir らは単語分散表現の類似度に基づく尺度を新たに提案した [8]．2 つの対話コーパスを用いた分析実験の結果，bag-of-words に基づくエントレイメント尺度と比べ，単語分散表現に基づく尺度が対人行動情報を把握するために有用であることが示された．なお Nasir らは深層学習による文脈埋め込み表現を用いた尺度の可能性に言及していた．その方向性の研究として，Bidirectional Encoder Representations from Transformers (BERT) [13] によって埋め込まれた文ベクトルの相関に基づくエントレイメント尺度が Liu らにより提案されている [14]．

本研究では，著者らが直近に提案した単語分散表現に基づくエントレイメントスコアを用いたリランキング手法に対して，新たに BERT の文脈埋め込みに基づくエントレインメントスコアを用いることを

提案する．そして，これらのスコアに基づくリランキング手法で得られた応答の自動評価を比較する．

2 エントレインメントスコアを用いた対話応答のリランキング

本研究では，著者らが直近で提案したリランキング手法 [11] を適用する．手法の概要を図 1 に示す．

図 1 リランキング手法の概要

2.1 ニューラル雑談対話モデル

ニューラル雑談対話モデルは，入力に t ターン目の発話文 U_t までの対話履歴 $\boldsymbol{U} = \{U_1, U_2 \ldots, U_t\}$ を受け取り[1)]，出力として $t+1$ ターン目の n 個の n-best 応答候補リスト $\boldsymbol{R} = [R_1, R_2 \ldots, R_n]$ および応答候補の尤度リスト $\boldsymbol{l_R} = [l_{R_1}, l_{R_2} \ldots, l_{R_n}]$ を出力する．ただし，応答候補リスト $\boldsymbol{R}$ は尤度の降順になっており，$l_{R_1} \geq l_{R_2} \geq \ldots, \geq l_{R_n}$ を満たす．

2.2 エントレインメントスコアの定義

本研究では，局所的な（固定窓幅における対話文脈の）言語的エントレインメントを測定するフレームワークである Local Interpersonal Distance (LID) [8] を用いてエントレインメントスコアを算出する．著者らの従来手法 [11] では，ある発話者による発話 U_t に関し，その対話相手へのエントレインメント度合い LID_{U_t} を次式により算出した．

$$\mathrm{LID}^{(\mathrm{WMD})}_{U_t} = \min_{U^{\mathrm{partner}} \in \boldsymbol{U}^{\mathrm{partner}} \subset \boldsymbol{U}} \mathrm{WMD}(U^{\mathrm{partner}}, U_t) \quad (1)$$

ここで，$\boldsymbol{U}^{\mathrm{partner}}$ は U_t より過去に観測された発話のうち，対話相手による過去 2 ターンの発話文集合を表す．また，$\mathrm{WMD}(U^{\mathrm{partner}}, U_t)$ は Word Mover's Distance (WMD) [16] を用いて計算される発話 $U^{\mathrm{partner}}, U_t$ 間の単語分散表現空間上での意味的距離である．$\mathrm{LID}^{(\mathrm{WMD})}_{U_t}$ は 0 以上の値をとる実数であり，値が小さいほど発話 U_t が対話相手の発話集合 $\boldsymbol{U}^{\mathrm{partner}}$ に対して同調しているとみなせる．

本稿では，式 (1) の代わりに，新たに BERTScore [17] に基づくエントレインメント度合い $\mathrm{LID}^{(\mathrm{BERT})}_{U_t}$ を用いることを提案する．BERTScore は単語ごとに独立して埋め込み表現を獲得するのではなく，Self-Attention により文全体の意味を考慮して埋め込み表現を獲得する．この文脈埋め込みを発話文間の類似度算出に用いることで，発話文全体を考慮したエントレインメントスコアを期待する．

$$\mathrm{LID}^{(\mathrm{BERT})}_{U_t} = \min_{U^{\mathrm{partner}} \in \boldsymbol{U}^{\mathrm{partner}} \subset \boldsymbol{U}} \{1 - \mathrm{BERTScore}(U^{\mathrm{partner}}, U_t)\} \quad (2)$$

$\mathrm{LID}^{(\mathrm{BERT})}_{U_t}$ は 0 以上 1 以下の値をとる実数であり，値が 0 に近いほど発話 U_t が対話相手の発話集合 $\boldsymbol{U}^{\mathrm{partner}}$ に対して同調しているとみなせる．

2.3 エントレインメント予測モデル

t ターン目までの対話履歴 $\boldsymbol{U}$ が入力されたとき，$t+1$ ターン目の発話文 U_{t+1} が持つべきエントレインメントスコア $\mathrm{LID}_{U_{t+1}}$ を予測するようモデルを学習する [12]．図 2 に示すように，Gated Reccurent Unit (GRU) [18] を用いた階層的エンコーダモデルと線形変換層（Affine）を用いたエントレインメントデコーダモデルで構成する．

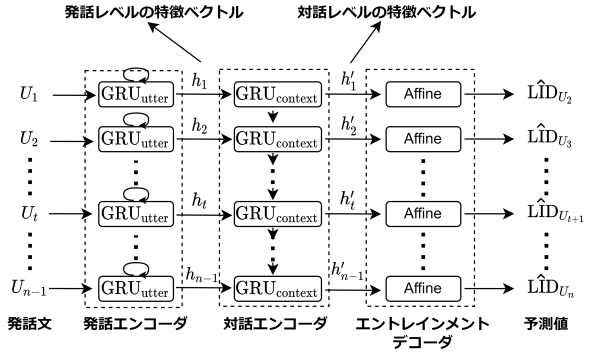

図 2 エントレインメント予測モデルの概要

発話文中の各単語 $u_{t,i} \in U_t$ を，式 (3) で定義するエンコーダで発話レベルの特徴ベクトル $\boldsymbol{h_t} = \boldsymbol{h_{t,|U_t|}}$ に変換する[2)]．

$$\boldsymbol{h_{t,i}} = \mathrm{GRU}_{\mathrm{utter}}(\boldsymbol{h_{t,i-1}}, \mathrm{Embedding}(u_{t,i})) \quad (3)$$

ここで，$\mathrm{Embedding}(\cdot)$ は，U_t における各単語 $u_{t,i} \in U_t$ を固定長の密ベクトル表現に変換する単語埋め込み層である．

次に，式 (4) で定義される対話エンコーダを用いて，各ターンまでに得られた発話レベルの特徴ベクトル $\boldsymbol{h_t}$ の系列を対話レベルの特徴ベクトル $\boldsymbol{h'_t}$ に統合する．

$$\boldsymbol{h'_t} = \mathrm{GRU}_{\mathrm{context}}(\boldsymbol{h_t}, \boldsymbol{h'_{t-1}}) \quad (4)$$

1) 文献 [15] の実験に従い，実際の入力には直近 4 発話の対話履歴 $\boldsymbol{U_t} = \{U_{t-3}, U_{t-2} \ldots, U_t\}$ を用いる．

2) $\boldsymbol{h_t}$ は発話エンコーダ $\mathrm{GRU}_{\mathrm{utter}}$ に対する最後の単語 $u_{t,|U_t|}$ の入力に対応する，隠れベクトル $\boldsymbol{h_{t,|U_t|}}$ である．

さらに，エントレインメントデコーダ Affine(·) を用いて，各ターンの特徴ベクトル $\boldsymbol{h}'_t$ からエントレインメントの予測値 $\hat{\mathrm{LID}}_{U_{t+1}}$ を得る．

$$\hat{\mathrm{LID}}_{U_{t+1}} = \mathrm{Affine}(\boldsymbol{h}'_t) \tag{5}$$

2.4 エントレインメントを用いたリランキング機構

まず，対話履歴からみて望ましいエントレインメントスコアを式 (5) で予測し，それを $\mathrm{LID}_{\mathrm{target}} = \hat{\mathrm{LID}}_{U_{t+1}}$ とおく．そして，ニューラル雑談対話モデルによる応答候補 $R_i \in \boldsymbol{R}$ が実際にもたらすエントレインメントスコア LID_{R_i} との差の絶対値を，エントレインメントの実現度 $d_{R_i,\mathrm{target}}$ として定義する．

$$d_{R_i,\mathrm{target}} = |\mathrm{LID}_{R_i} - \mathrm{LID}_{\mathrm{target}}| \tag{6}$$

著者らのこれまでの手法 [11] と同様に，実現度 $d_{R_i,\mathrm{target}}$ と応答尤度 l_{R_i} の両方をバランスよく考慮して応答候補リスト $\boldsymbol{R}$ 内をリランキングする．具体的には，$R_i \in \boldsymbol{R}$ に対し，正規化した尤度 l_{R_i} と正規化した実現度 $d_{R_i,\mathrm{target}}$ の F-beta を算出し，最大値をとる応答 $R_{\text{F-beta}}(\beta)$ を選択する．

$$R_{\text{F-beta}}(\beta) = \arg\max_{R_i \in \boldsymbol{R}} \left[(1+\beta^2) \frac{\widetilde{\frac{1}{|l_{R_i}|}} \times \widetilde{\frac{1}{d_{R_i,\mathrm{target}}}}}{\beta^2 \widetilde{\frac{1}{|l_{R_i}|}} + \widetilde{\frac{1}{d_{R_i,\mathrm{target}}}}} \right] \tag{7}$$

上式において，$\tilde{\cdot}$ はそれぞれの応答候補リストを基準とした Min-Max 正規化を表し，β は 0 以上の実数となる重み係数である．β が 1 より大きいほどエントレインメントを重視して応答を選択する．本研究では，$\beta = 1$ と設定した．

3 評価実験

本章では，ニューラル雑談対話モデルの応答リランキングに提案手法を適用した場合の応答の自然性を自動評価指標を用いて評価する．

3.1 データセット

評価実験では，感情的な状況下での共感的な対話を収録した雑談対話コーパス JEmpatheticDialogues [15] を用いた．コーパス内の対話数は 20,000，1 対話は 4 ターンである．文献 [15] の実験に従い，16,667 対話（83.34%）を学習データ，1,667 対話を検証データ（8.34%），1,666 対話（8.33%）を評価データとした．

3.2 実験設定

n-best 応答候補リストを生成するニューラル雑談対話モデルとして，JEmpatheticDialogues でファインチューニングされた Transformer Encoder-Decoder モデル [15] を用いた．n-best 応答候補リストのサイズは $n = 80$ に設定した．本実験では，次に示すベースライン応答とリランキング応答を自動評価によって比較する．

- R_1（ベースライン，尤度最大の応答）
- $R_{(\mathrm{BERT})}$（提案手法，$\mathrm{LID}^{(\mathrm{BERT})}_{U_t}$ に基づく応答）
- $R_{(\mathrm{WMD})}$（従来手法，$\mathrm{LID}^{(\mathrm{WMD})}_{U_t}$ に基づく応答）

また，エントレインメント予測モデルの教師スコアを用いて，式 (6) について $\mathrm{LID}_{\mathrm{target}} = \mathrm{LID}_{U_{t+1}}$ として設定した場合のリランキング応答も比較する．これはテストデータ中の対話履歴に存在する実際の応答発話のエントレインメントスコアを目標とする設定であり，エントレインメント予測モデルの誤差が無い場合に相当する．

エントレインメントスコアの計算に使用する単語分散表現モデルについては，Akama らが提案したスタイルの類似性と統語的・語義的な類似性の両方を考慮した単語分散表現モデル [19] を用いた．単語分散表現モデルの学習には，学習データセット内に含まれる発話を用いた．

エントレインメント予測モデルにおけるハイパーパラメータは，単語埋め込み層の次元を 300，GRU の中間層の次元を 300，層数を 1 に設定した．また，使用する語彙サイズは 8,000 とし，未知語は特殊記号 “UNK” に置き換えた．予測モデルの訓練においてはバッチサイズを 64，学習率を 1×10^{-3} として，Optimizer には Adam[20] を使用した．

3.3 評価指標

システム応答 $R_\#$ と対話履歴に対する実際の応答発話 U_{t+1} とのエントレインメント誤差を評価するために，次式で表される symmetric mean absolute percentage error (sMAPE) を用いた．

$$\begin{aligned} &\mathrm{sMAPE}(\mathbf{LID}_{R_\#}, \mathbf{LID}_{U_{t+1}}) \\ &= \frac{200}{n} \sum_{j=1}^{n} \frac{|\mathrm{LID}_{R_\#,j} - \mathrm{LID}_{U_{t+1},j}|}{|\mathrm{LID}_{R_\#,j}| + |\mathrm{LID}_{U_{t+1},j}|} \end{aligned} \tag{8}$$

ここで n は評価データの数である．また LID には $\mathrm{LID}^{(\mathrm{BERT})}$ と $\mathrm{LID}^{(\mathrm{WMD})}$ の 2 種類がありエントレインメント誤差はそれぞれの尺度について算出する．

表 1 自動評価結果（$\mathrm{LID}_{\mathrm{target}} = \hat{\mathrm{LID}}_{U_{t+1}}$）

手法	sMAPE(BERT)	sMAPE(WMD)	BLEU-1	BLEU-2	Distinct-1	Distinct-2	Embedding average
ベースライン	14.28%	20.84%	23.9	7.2	0.891	0.906	0.435
BERTScore	**9.71**%	19.08%	23.1	6.4	0.886	0.911	0.428
WMD	12.16%	**14.72**%	23.2	6.6	0.907	0.887	0.430

表 2 自動評価結果（$\mathrm{LID}_{\mathrm{target}} = \mathrm{LID}_{U_{t+1}}$）

手法	sMAPE(BERT)	sMAPE(WMD)	BLEU-1	BLEU-2	Distinct-1	Distinct-2	Embedding average
BERTScore	**2.92**%	18.03%	23.9	7.1	0.885	0.910	0.441
WMD	11.85%	**4.29**%	24.5	7.2	0.883	0.907	0.453

$\mathrm{LID}_{U_{t+1}}$ はテストデータにおける対話履歴に対する実際の応答発話のエントレインメントスコアである．応答の自動評価指標には BLEU [21]，Distinct [22]，Embedding Average [23] を用いた．

3.4 実験結果

表 1 に予測モデル（$\mathrm{LID}_{\mathrm{target}} = \hat{\mathrm{LID}}_{U_{t+1}}$ とした場合）によるリランキング応答の自動評価の結果を示す．これは提案手法と従来手法による評価実験の結果である．BERTScore と WMD のどちらの手法においても，ベースラインに比べてリランキングされた応答はそれぞれの尺度に対応するエントレイメント誤差（sMAPE）が減少した．さらに，リランキングで用いなかった尺度に対応するエントレイメント誤差も僅かに減少する．これは BERTScore と WMD が埋め込み表現を獲得する際に，一部は同じ言語特徴量を捉えている可能性がある．一方で，どちらの尺度を用いた場合でも自動評価の結果に大きな差はみられなかった．特に，埋め込み表現を用いたにも関わらず，Embedding Average は向上しなかった．

次に，表 2 に教師スコア（$\mathrm{LID}_{\mathrm{target}} = \mathrm{LID}_{U_{t+1}}$ とした場合）によるリランキング応答の自動評価の結果を示す．これはエントレインメント予測モデルの誤差が無い場合に相当する．表 1 と比較すると，どちらの手法においてもベースラインに比べてリランキングされた応答はそれぞれの尺度に対応するエントレイメント誤差が大幅に減少した．しかし，リランキングで用いなかった尺度に対応するエントレイメント誤差はそれほど減少しない．このことから BERTScore と WMD は異なる言語的特徴を捉えている可能性がある．自動評価の結果については大きな差はみられない．しかし，表 1 と比較すると全体として自動評価の結果が僅かに良い傾向にある．このことから，エントレインメント予測モデルの精度が向上することで応答の自然性などが向上する可能性がある．ただし，雑談対話システムに対する BLEU などの自動評価指標は人間による主観評価とはほとんど相関がないことが報告されており [23]，主観評価は異なる結果が得られる可能性がある．

4 おわりに

本研究では，著者らが直近に提案した単語分散表現に基づくエントレイメントスコアを用いたリランキング手法に対して，新たに BERT の文脈埋め込みに基づくエントレイメントスコアを導入した．また，これらのエントレイメントスコアに基づく応答を自動評価し，性能を比較した．従来手法と提案手法によるリランキングはどちらもベースラインに比べてエントレイメント誤差が減少した．一方，応答の自動評価の結果は提案手法と従来手法の間でほとんど違いがみられなかった．

今後は，BERTScore に基づくリランキング結果について主観評価を実施する．また，異なるエントレイメントスコアが捉える言語的側面（語彙，統語構造，文体，ターンテイキング，話者の意図など）を分析し，それらを網羅的に捉える新たなエントレイメントスコアの算出方法を検討する．

謝辞

本研究は，JSPS 科研費 22K17958, 22H04873, 20H04484 の助成を受けた．

参考文献

[1] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, Vol. 22, No. 6, p. 1482, 1996.

[2] David Reitter and Johanna D Moore. Predicting success in dialogue. 2007.

[3] Kate G Niederhoffer and James W Pennebaker. Linguistic style matching in social interaction. **Journal of Language and Social Psychology**, Vol. 21, No. 4, pp. 337–360, 2002.

[4] Arthur Ward and Diane Litman. Measuring convergence and priming in tutorial dialog. **University of Pittsburgh**,

2007.

[5] Nick Campbell and Stefan Scherer. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In **Eleventh Annual Conference of the International Speech Communication Association**, 2010.

[6] Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David Traum, and Satoshi Nakamura. Analyzing the effect of entrainment on dialogue acts. In **Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, September 2016.

[7] Ani Nenkova, Agustin Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In **Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers**, pp. 169–172. Association for Computational Linguistics, 2008.

[8] Md. Nasir, Sandeep Nallan Chakravarthula, Brian R.W. Baucom, David C. Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. Modeling Interpersonal Linguistic Coordination in Conversations Using Word Mover' s Distance. In **Proc. INTERSPEECH 2019**, pp. 1423–1427, 2019.

[9] 水上雅博, 吉野幸一郎, Graham Neubig, 中村哲. エントレインメント分析に基づく応答文選択モデルの評価. 言語処理学会第 23 回年次大会 (NLP2017), 茨城, 3 2017.

[10] Seiya Kawano, Masahiro Mizukami, Koichiro Yoshino, and Satoshi Nakamura. Entrainable neural conversation model based on reinforcement learning. **IEEE Access**, Vol. 8, pp. 178283–178294, 2020.

[11] 金崎翔大, 河野誠也, 湯口彰重, 桂井麻里衣, 吉野幸一郎. エントレインメント予測に基づいたニューラル雑談対話モデルの応答リランキング. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 3Yin248–3Yin248, 2022.

[12] 金崎翔大, 河野誠也, 湯口彰重, 桂井麻里衣, 吉野幸一郎. 対話における後続発話のエントレインメント予測. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 50–55, 2021.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[14] Yuning Liu, Aijun Li, Jianwu Dang, and Di Zhou. Semantic and acoustic-prosodic entrainment of dialogues in service scenarios. In **Companion Publication of the 2021 International Conference on Multimodal Interaction**, ICMI '21 Companion, p. 71–74, New York, NY, USA, 2021. Association for Computing Machinery.

[15] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[16] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In **International Conference on Machine Learning**, pp. 957–966, 2015.

[17] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[18] 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 **Proc. of EMNLP**, pp. 1724–1734, 2014.

[19] Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. Unsupervised learning of style-sensitive word vectors. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 572–578, 2018.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

[21] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[23] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics.

付録（Appendix）

A　実際の対話例

本節には，評価実験における実際の対話履歴とリランキング応答について 3 つの具体例を示す．各表では対話者 A，B による対話履歴を示した後に，対話履歴に続く実際の応答，R_1（ベースライン応答，応答尤度最大），$R_{(\mathrm{BERT})}$（提案手法によるリランキング応答），$R_{(\mathrm{WMD})}$（従来手法によるリランキング応答）を順に示す．

A.1　対話例 1

$R_{(\mathrm{BERT})}$ には 5 番目の応答候補が選択され，$R_{(\mathrm{WMD})}$ には 7 番目の応答候補が選択された．

表 3　対話例 1

A	今日は、仕事が終わった後に合コンの予定が入ってるんだ。
B	楽しみだね！
実際の応答	うん、夕方から、もう、ワクワクそわそわしてるよ。
R_1	うん、今から、わくわくしてるんだ。
$R_{(\mathrm{BERT})} = R_5$	ずっと前から楽しみにしてたから、早く家に帰りたくてうずうずしてる。
$R_{(\mathrm{WMD})} = R_7$	うん、合コンって初めてだから緊張するけど、楽しいと良いな。

A.2　対話例 2

$R_{(\mathrm{BERT})}$ には 1 番目の応答候補が選択された．つまり BERTScore に基づく尺度を用いたリランキング手法を適用しても，リランキングする前と同じ応答が出力された．$R_{(\mathrm{WMD})}$ には 3 番目の応答候補が選択された．

表 4　対話例 2

A	中学校転校の日に、教室に入ると黒板にメッセージが書いてあったよ。
B	それうれしいね。みんなと離れるのがいやになっちゃうよね。
実際の応答	全員から一言ずつメッセージが書いてあって感動したよ。転校さみしいな。
R_1	うん。今でも黒板にメッセージを見ると、青春を思い出すよ。
$R_{(\mathrm{BERT})} = R_1$	うん。今でも黒板にメッセージを見ると、青春を思い出すよ。
$R_{(\mathrm{WMD})} = R_3$	うん。あの日はなんだか思い出が蘇ってきて、感傷的になってしまったよ。

A.3　対話例 3

$R_{(\mathrm{BERT})}$ には 8 番目の応答候補が選択され，$R_{(\mathrm{WMD})}$ には 2 番目の応答候補が選択された．

表 5　対話例 3

A	リニア新幹線って大丈夫なのかな？
B	たまにニュースで環境への影響が話題になってるよね
A	山を貫通させるわけでしょ？土砂とか水とか不安なんだけど
実際の応答	しっかり調査して進めてほしいよね
R_1	そうだよね。リニアが開通する頃にはだいぶ復旧してると思うから、大丈夫だよ
$R_{(\mathrm{BERT})} = R_8$	と、思うけどどうなんだろうね？とりあえず安全には配慮してるとは思うけど。
$R_{(\mathrm{WMD})} = R_2$	確かにそうだよね。大きな影響が出なきゃいいんだけど