ゼロ照応解析に基づく項省略補完を取り入れた対話応答生成

上山彩夏　狩野芳伸
静岡大学大学院総合科学技術研究科
aueyama@kanolab.net　kano@inf.shizuoka.ac.jp

概要

人間同士の対話では，言外の情報が重要である．昨今の対話システムには，対話履歴から応答を直接生成する手法が広く用いられているが，対話履歴内の述語の項は頻繁に省略されるため，統計的なパターンを学習するのみでは，言葉には表れない発話意図を汲み取ることが難しい．そこで，対話履歴内の省略された情報を推測し，明示的に補完した対話履歴から応答を生成する Dialogue Completion using Zero Anaphora Resolution framework (DCZAR) を提案する．自動評価及び人手評価の結果，DCZAR が応答の首尾一貫性と魅力度を大幅に向上させることを確認した．

1　はじめに

人間同士の対話は，常識的知識や共有認識からなる共通基盤に依存している [1, 2]．表 1 の対話では，多くの情報が省略されているが，2 人の話者は常識的知識や文脈をもとに相手の意図を汲み取ることができている．昨今の対話システムは深層学習技術の発展により，着実に性能が改善されているものの [3, 4, 5]，表 1 のような発話の意図を正確に捉えることはできていない [6, 7]．これは，人間とシステムの間に共通基盤が存在しないことに起因すると考えられる．既存の対話システムは，対話履歴から応答を直接生成する手法が広く用いられており，共通基盤構築の過程を明示的に扱っていない．また，対話履歴には多くの省略があるため，発話の意図を学習することが困難となっている．この問題に対し，外部知識を導入し，人間の常識的知識をモデルに与えることで，人間とモデルのギャップを埋める知識ベース (KB) を用いる手法が数多く提案されているが [8, 9, 10]，KB の構築にかかるコストは大きく，異なるドメインやモデルへの転用も容易ではない．

表 1　対話例．highlight は省略された情報．

話者 A:	最近、友人が学校に来ていません。
	私は 友人に 連絡して良いですかね？
話者 B:	友人が 1 人で悩んでいるかもしれない
	ので、 あなたは 友人に 連絡すべきです。

本研究では，外部知識を利用せずに，発話の意図を捉える手法を検討する．対話相手の発話意図を汲み取るためには，登場する人物や事物の役割とその変遷を把握することが重要である．そのため，それら人物や事物とその役割を推測し，明示的に取り入れることで，相手の意図に沿った応答を生成できるようになり，首尾一貫した対話が実現できると期待される．そこで，省略された照応を推測するゼロ照応解析の考えから着想を得て，対話履歴内の省略された情報を推測し，明示的に補完した対話履歴から応答を生成する**ゼロ照応解析に基づく対話応答生成フレームワーク** (Dialogue Completion using Zero Anaphora Resolution framework; **DCZAR**) を提案する．DCZAR は，**述語項構造解析** (Predicate Argument Structure analysis; **PAS**) モデル，**対話補完** (Dialogue Completion; **DC**) モデル，**応答生成** (Response Generation; **RG**) モデルの 3 つのモデルで構成する．PAS モデルは対話履歴内の省略された項を解析し，DC モデルは補完する項と補完位置を決定し，明示的に補完する．その後，RG モデルは補完された対話履歴から応答を生成する．PAS モデル，RG モデルは事前学習済みモデルを各タスクに対応するデータセットで fine-tuning したモデルで，DC モデルは事前学習済みモデルを fine-tuning せずに利用する．本研究では，2 種類の事前学習タスク (Cloze [11]，PZERO [12]) と 2 種類のデータセット (Wikipedia，Twitter) を組み合わせ，計 4 種類の事前学習済みモデルを構築し，各タスクにおいてどのモデルが優れた性能を達成するか検証した．

本研究の主要な貢献は以下の三点である．

- ゼロ照応解析に基づく項省略補完を応答生成モデルに取り入れることで，応答の首尾一貫性と魅力度が大幅に向上することを示した (4.5 節)．
- 対話データに類似した特徴を持つ Twitter データにより事前学習することで，ゼロ照応解析の性能を向上させた (4.3 節)．
- 対話補完モデルが対話に含まれる省略を十分な性能で補完できることを確認した (4.4 節)．

2 関連研究

2.1 対話応答生成

対話応答生成は，対話履歴に続く応答を生成するタスクであり，ソース文からターゲット文を生成する系列変換問題として定式化できる [7, 13]．

2.2 ゼロ照応解析

ゼロ照応解析は，与えられた文章に対し，述語の省略された引数 (ガ格，ヲ格，ニ格) を検出し，その先行詞を同定するタスクであり，述語項構造解析タスクの一部として定式化できる．ゼロ照応は述語とその項の位置関係に応じて，文内ゼロ[1] (*intra*)，文間ゼロ[2] (*inter*)，外界ゼロ[3] (*exo*) に分類される [14]．また，述語の引数が述語と直接係り受け関係にある場合，その引数は構文従属引数 (*dep*) である．

日本語ゼロ照応解析に関して，様々な研究が行われてきた [15, 16, 17]．Konno ら [12] は，ゼロ照応解析にはゼロ代名詞と先行詞の文脈的なつながりを理解するための常識的な知識が重要だと考え，以下の事前学習タスクと fine-tuning 手法を提案した．

擬似ゼロ代名詞解析 (PZERO) 入力系列に 2 回以上出現する名詞句のうち 1 つを `[MASK]` に置換し，`[MASK]` に埋まるべき名詞句を入力系列 X から選択する事前学習タスクである．`[MASK]` を系列 X から選択するタスクは，ゼロ代名詞に対応する先行詞を同定するタスクと類似しているため，モデルがゼロ照応解析に必要な知識を獲得できると期待される．

類似ゼロ代名詞に基づく項選択 (AS-PZERO) PZERO で訓練されたパラメータを用いて，PZERO と同様の形式で述語項を解析する手法である．モデルは，系列 X 及び X に含まれる述語を入力として受け取り，述語の項となる単語を X から選択する．述語の項が入力系列に存在しない場合は，モデルに `[CLS]` を選択させ，項をさらに 4 つのカテゴリ[4] (`author`,`reader`,`general`,`none`) へ分類する．

3 提案手法

我々が提案する DCZAR の概要を図 1 に示す．

PAS モデル 長さ T の対話履歴 $X = \{x_1, x_2, \ldots, x_T\}$ に対して，述語項構造解析を行い，X に含まれる n 個の述語 $P = \{p_1, p_2, \ldots, p_n\}$ の l 格[5] の項 $A_l = \{a_{l,1}, a_{l,2}, \ldots, a_{l,n}\}$ を予測する．

DC モデル 対話履歴 X と述語 P，PAS モデルが予測した述語の項 A_l を用いて，X 内の省略を明示的に補完する．補完を行う際には，補完する項とその補完位置を決定する必要がある．ある述語 p_i とその直前の述語 p_{i-1} の間 (探索範囲 r_i) に項 $a_{l,i}$ の先行詞が出現しているかを確認し，現れていなければ，その項は補完の対象とする．その後，項 $a_{l,i}$ を r_i 内の各トークン間に挿入する場合と項 $a_{l,i}$ を挿入しない場合の文を作成し，文章の自然さを表す指標である Pseudo-log-likelihood scores (PLLs) [18] を算出する．そして，最もスコアが高い文を適切な位置に補完された文とみなし，RG モデルの入力に利用する．

RG モデル DC モデルに補完された対話履歴とそれに続く応答により訓練されたモデルであり，推論の際は対話履歴のみを入力とし，応答生成を行う．

4 実験

4.1 事前学習

2 種類のタスク (Cloze [11]，PZERO [12]) と 2 種類の日本語データセット (Wikipedia，Twitter) を組み合わせ，計 4 種類の事前学習済みモデル (wiki-cloze，twitter-cloze，wiki-pzero，twitter-pzero) を構築した．また，モデルの初期パラメータとして，`bert-base-japanese-whole-word-masking` を用いた．

4.2 比較モデル

4.1 節に示した事前学習済みモデルを PAS，DC，RG の各モデルに使用し，性能比較を行った．比較パターンを表 2 に示す．例えば，(e) の RG`wiki-cloze`

1) 先行詞が同一文中に存在する．
2) 先行詞が述語の文より前方に存在する．
3) 先行詞が文中に存在しない．
4) `author`,`reader`,`general` は外界ゼロの細分類，`none` は項が存在しないことを表す．
5) l はガ，ヲ，ニのいずれかを表す．

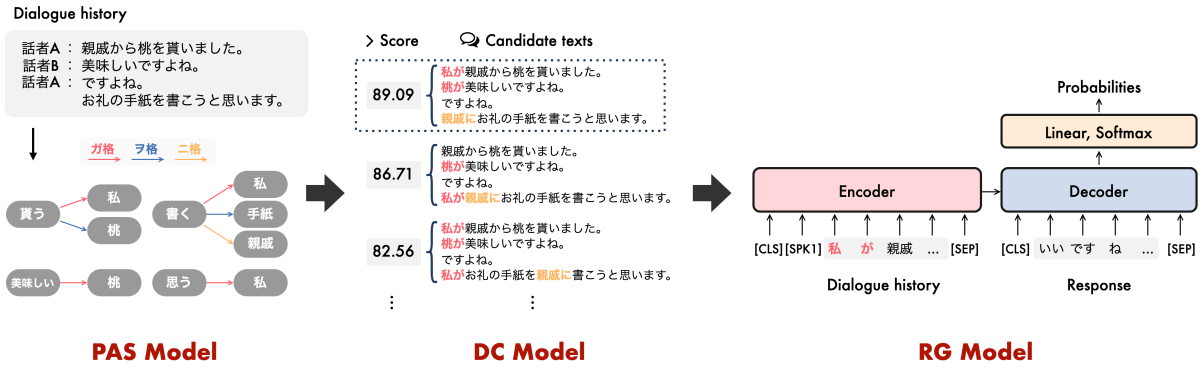

図 1　DCZAR の概要．PAS モデルは対話履歴内の省略された項を解析し，DC モデルは補完する項と補完位置を決定し，省略を明示的に補完する．RG モデルは補完された対話履歴から応答を生成する．

表 2　PAS，DC，RG モデルの比較パターン

ID	PAS Model	DC Model	RG Model
(a)	N/A	N/A	wiki-cloze
(b)	N/A	N/A	twitter-cloze
(c)	N/A	N/A	wiki-pzero
(d)	N/A	N/A	twitter-pzero
(e)	wiki-cloze	wiki-cloze	wiki-cloze
(f)	twitter-cloze	twitter-cloze	twitter-cloze
(g)	wiki-pzero	wiki-cloze	wiki-pzero
(h)	twitter-pzero	twitter-cloze	twitter-pzero

表 3　PAS モデルの自動評価結果 (F_1)

ID	Model	ZAR				*dep*	All
		All	*intra*	*inter*	*exo*		
(e)	PAS$_{\texttt{wiki-cloze}}$	62.27	68.39	**44.63**	67.77	**94.17**	83.67
(f)	PAS$_{\texttt{twitter-cloze}}$	62.21	68.04	40.68	70.34	94.15	83.73
(g)	PAS$_{\texttt{wiki-pzero}}$	62.68	68.35	43.02	69.99	93.96	83.75
(h)	**PAS$_{\texttt{twitter-pzero}}$**	**63.25**	**68.68**	42.07	**72.04**	93.81	**83.87**

は前処理として，PAS$_{\texttt{wiki-cloze}}$ と DC$_{\texttt{wiki-cloze}}$ を用いる．また，(a) から (d) は補完を一切行わないベースラインモデルであり，(h) は提案する事前学習済みモデル (twitter-pzero) のみを組み合わせたモデルである．なお，DC モデルで利用する PLLs は Cloze タスクが解けるモデルを必要とするため，(g) と (h) の DC モデルは，wiki-pzero と twitter-pzero の代わりに wiki-cloze と twitter-cloze をそれぞれ適用している．

4.3　実験 1：PAS モデル単体の性能評価

本実験では，表 2 に示した (e) から (h) の PAS モデルの述語項構造解析の性能 (F_1) を評価した．PAS モデルは事前学習済みモデルに対し，NAIST Text Corpus 1.5 [14] を用いて，AS-PZERO による fine-tuning を行ったモデルである．PAS モデルへの入力は，述語を含む文とその前方文で，述語に対応する先行詞と格情報を出力するように学習される．

表 4　DC モデルの人手評価結果

ID	Model	適切さの評価
(e)	DC$_{\texttt{wiki-cloze}}$	74.80% (187 / 250)
(f)	DC$_{\texttt{twitter-cloze}}$	77.20% (193 / 250)
(g)	DC$_{\texttt{wiki-pzero}}$	72.40% (181 / 250)
(h)	**DC$_{\texttt{twitter-pzero}}$**	**84.80%** (**212** / 250)

実験結果を表 3 に示す．提案した PAS$_{\texttt{twitter-pzero}}$ がゼロ照応解析において，最も高い性能を達成した．Twitter データで事前学習したモデルは，Wikipedia データで事前学習したモデルと比較して性能が高く，特に外界ゼロ (*exo*) で大きな向上を示している．

4.4　実験 2：DC モデル単体の性能評価

本実験では，表 2 に示した (e) から (h) の DC モデルの補完性能を人手評価により評価した．人手評価には，JPersonaChat と JEmpatheticDialogues [19] から 4 つのモデルごとにランダムサンプリングした各 250 件を用いた．5 人の評価者に対し，補完前の対話履歴と補完後の対話履歴を提示し，補完した項と補完位置の双方が適切かを決定してもらった．

表 4 に実験結果を示す．提案した DC$_{\texttt{twitter-pzero}}$ が対話補完において，最も性能が高かった．Cloze，PZERO ともに，Wikipedia データではなく，Twitter データを事前学習に用いることで，性能が向上している (74.80 → 77.20，72.40 → 84.80)．また，表 3 においても，(h) が最も優れた性能を示していることから，対話補完の性能は述語項構造解析の性能に関連していると考えられる．

4.5　実験 3：DCZAR の性能評価

本実験では，表 2 に示した (a) から (h) の対話応答生成の性能を自動評価及び人手評価により評価し

表 5　RG モデルの自動評価結果

ID	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE-L	DIST 1	DIST 2	BERT Score
(a)	RG$_{\text{wiki-cloze}}$	25.29	6.14	2.02	0.69	9.57	**12.20**	**29.32**	69.70
(e)	+ **DCZAR (ours)**	24.50	5.65	1.78	0.55	**14.50**	11.72	28.50	69.45
(b)	RG$_{\text{twitter-cloze}}$	25.65	**6.50**	2.10	0.70	9.79	12.04	29.02	69.84
(f)	+ **DCZAR (ours)**	**25.72**	6.16	1.96	0.66	11.65	12.14	28.95	69.73
(c)	RG$_{\text{wiki-pzero}}$	25.59	6.31	2.09	0.72	13.72	12.09	28.96	**69.90**
(g)	+ **DCZAR (ours)**	25.45	6.08	1.96	0.63	6.49	12.06	29.14	69.75
(d)	RG$_{\text{twitter-pzero}}$	25.00	6.08	2.02	0.69	11.99	12.17	29.31	69.74
(h)	+ **DCZAR (ours)**	25.50	6.17	**2.11**	**0.73**	9.41	11.72	28.54	69.77

表 6　RG モデルの人手評価結果．*/**は，カイ二乗検定で $p < 0.05/0.01$ での統計的有意差があることを示す．N/A は，多数決で「分からない」が選択された場合と 3 人の評価者で評価が割れた場合の件数である．

ID	Model	文法的流暢性	首尾一貫性	魅力度
(a)	RG$_{\text{wiki-cloze}}$	**30**	45	44
(e)	+ **DCZAR (ours)**	28	**54****	**55****
	N/A	42	1	1
(b)	RG$_{\text{twitter-cloze}}$	30	43	46
(f)	+ **DCZAR (ours)**	**34**	**57****	**54****
	N/A	36	0	0
(c)	RG$_{\text{wiki-pzero}}$	**34**	45	**51**
(g)	+ **DCZAR (ours)**	33	**52****	47
	N/A	33	3	2
(d)	RG$_{\text{twitter-pzero}}$	32	38	41
(h)	+ **DCZAR (ours)**	**38**	**62****	**59****
	N/A	30	0	0

た．RG モデルは，BERT を Encoder 及び Decoder に利用する BERT2BERT [20] に対し，JPersonaChat と JEmpatheticDialogues を用いて，fine-tuning を行ったモデルである．(a) から (d) はベースラインモデル，(e) から (h) は DCZAR を適用した提案モデルであり，ベースラインモデルの入力には，データセットに含まれる対話履歴 H をそのまま利用し，提案モデルの入力には，H 内の省略を補完した対話履歴 H' を用いた．

自動評価には，BLEU [21]，ROUGE-L [22]，DIST-N [23]，BERTScore [24] を採用した．人手評価は，JPersonaChat と JEmpatheticDialogues から，4 種類の事前学習済みモデルごとにランダムサンプリングした 100 件，計 400 件を全ての評価者が全件評価した．3 人の評価者に対し，対話履歴とベースラインモデル，提案モデルが生成した 2 つの応答を提示し，異なる評価基準 (文法的流暢性，首尾一貫性，魅力度) に基づいて 1 つを選択するか，「分からない」を選択するペアワイズ比較を行った．最終的な評価値は，3 名の多数決によって決定した．

表 5 に自動評価結果を示す．BLEU-1, 3, 4 及び ROUGE-L においては，提案モデルがベースラインモデルを上回ることを確認したが，有意な差は見られなかった．提案手法は，項を補うことによって，より首尾一貫した応答が生成できると期待されるが，その貢献は単語統計の結果には変化をもたらさないと考えられるため，これら自動評価尺度は必ずしも適切な評価尺度とは言えない．

表 6 に人手評価結果を示す．文法的流暢性は，全てのモデルにおいて有意差が認められなかったが，RG$_{\text{twitter-cloze}}$+DCZAR と RG$_{\text{twitter-pzero}}$+DCZAR はベースラインモデルの性能を上回った．有意差が認められなかった原因は，他の観点と比較して，N/A の件数が多いことが考えられる．首尾一貫性は，全ての提案モデルがベースラインモデルと比較して，有意に向上している．これは，応答を生成する際に，省略を明示的に補完した対話履歴を用いることが首尾一貫性の評価に寄与することを示している．魅力度は，RG$_{\text{wiki-pzero}}$+DCZAR 以外の全てのモデルにおいて，ベースラインモデルと比較して，有意に向上している．人手評価の結果より，DCZAR はより首尾一貫した，魅力的な応答の生成に貢献することが示された．

5　おわりに

本研究では，対話履歴内の省略された情報を推測し，明示的に補完した対話履歴から応答を生成する DCZAR を提案した．実験の結果，DCZAR を適用することで，より首尾一貫した，魅力的な応答を生成できることが分かった．今後の展望として，異なるタスクや他の言語への適用などが挙げられる．

謝辞

本研究は JSPS 科研費 JP22H00804，JP21K18115 の助成及び JST AIP 加速課題 JPMJCR22U4 の支援を受けたものです.

参考文献

[1] Robert C. Stalnaker. Assertion. pp. 315–332, 1978.

[2] Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. Vol. 13, No. 2, pp. 259–294, 1989.

[3] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**, pp. 8460–8478, 2022.

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and et al. Improving language models by retrieving from trillions of tokens. In **Proceedings of the 39th International Conference on Machine Learning**, pp. 2206–2240, 2022.

[5] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, and et al. LaMDA: Language models for dialog applications. arXiv:2201.08239, 2022.

[6] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, pp. 654–664, 2017.

[7] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 31, 2017.

[8] Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. Knowledge-grounded dialogue generation with pre-trained language models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 3377–3390, 2020.

[9] Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. Multi-sentence knowledge selection in open-domain dialogue. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 76–86, 2021.

[10] Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. Retrieval-free knowledge-grounded dialogue response generation with adapters. In **Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering**, pp. 93–107, 2022.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, 2019.

[12] Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. Pseudo zero pronoun resolution improves zero anaphora resolution. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3790–3806, 2021.

[13] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**, pp. 1237–1252, 2022.

[14] Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. Annotating predicate-argument relations and anaphoric relations: Findings from the building of the naist text corpus. Vol. 17, No. 2, pp. 25–50, 2010.

[15] Ryohei Sasano and Sadao Kurohashi. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In **Proceedings of 5th International Joint Conference on Natural Language Processing**, pp. 758–766, 2011.

[16] Souta Yamashiro, Hitoshi Nishikawa, and Takenobu Tokunaga. Neural Japanese zero anaphora resolution using smoothed large-scale case frames with word embedding. In **Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation**, 2018.

[17] Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 1920–1934, 2021.

[18] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2699–2712, 2020.

[19] 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. arXiv:2109.05217, 2021.

[20] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 264–280, 2020.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[22] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In **Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics**, pp. 605–612, 2004.

[23] 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**, pp. 110–119, 2016.

[24] 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.

[25] Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A Japanese predicate argument structure analysis using decision lists. In **Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing**, pp. 523–532, 2008.

[26] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**, pp. 2204–2213, 2018-07.

[27] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5370–5381, 2019.

A　データセットの統計情報

Wikipedia　日本語 Wikipedia のダンプデータ[6]を訓練/検証セットに含まれる事例数が 15M/3k(トークン数は 763M/220k) になるように分割した．本データを PZERO に用いる際には，名詞句を同定する必要があるが，Konno ら [12] と同様の手法で行った．

Twitter　Twitter API[7]を用いて収集したツイートを訓練/検証セットに含まれる事例数が 70M/30k(トークン数は 504M/200k) になるように分割した．名詞句の同定は，Wikipedia と同様の手法を採用した．

NAIST Text Corpus 1.5　述語と表層格の関係などが付与されたコーパスである．Taira ら [25] の手法に従って，訓練/検証/評価セットに分割した．

JPersonaChat　PersonaChat [26] の日本語版で，Sugiyama ら [19] の手法に従って，対話履歴と応答のペアを作成し，訓練/検証/評価セットに含まれる対話数が 50k/3k/4k になるように分割した．なお，本研究ではペルソナ記述文は利用していない．

JEmpatheticDialogues　EmpatheticDialogues [27] の日本語版で，Sugiyama ら [19] の手法に従って，対話履歴と応答のペアを作成し，訓練/検証/評価セットに含まれる対話数が 50k/3k/7k になるように分割した．

B　生成応答例の分析

表 7 に生成した応答例を示す．

文法的流暢性で評価に差が無かったのはなぜか？　文法的流暢性[8]において，有意差が認められなかったのは，他の観点と比較し，N/A が多いことに起因すると考えられる．対話例 1 は 3 人の評価者が「分からない」を選択して，N/A に分類された例である．この例では，2 つのモデルが異なる内容の応答を生成しているものの，どちらの応答も文法的には誤っておらず，このような場合には判断が難しいと考えられる．

提案手法は首尾一貫性の向上に寄与しているか？　対話例 2 は，提案モデルの応答がより首尾一貫[9]していると評価された例の 1 つである．対話例 2 の対話履歴には，「夫」，「ご主人」，「妻」，「専業主婦」といった人物の役割を表す名詞が多く存在しており，さらに「贈る」に対応する項が省略されているため，誰に関する話をしているのかが判断しづらい．RG twitter-pzero は妻に関する話だと誤った解釈をして，応答を生成しているが，提案した RG twitter-pzero+DCZAR は夫に関する話だと正しく解釈をして，応答を生成している．これは，省略された項を明示的に補完することが首尾一貫性の向上に寄与していることを示唆している．

表 7　生成応答例．highlight は提案モデルにより補完された項を表す．

対話例 1	
話者 A:	親の老後が不安。
話者 B:	この世代になると、 老後が 不安になってくるよね。
話者 A:	そうなんだよね。離れて暮らしてるし、 親の意向も分からないしでね。
RG wiki-cloze:	心配だよね。
+ **DCZAR (ours)**:	不安になるよね、何か考えておかないとね。
対話例 2	
話者 A:	専業主婦をしてると、 夫に感謝する場面がとても多いわ。
話者 B:	きっと優しいご主人なんだろうね。
話者 A:	お歳暮も、夫の親戚と同じように、 妻側の親戚にも嫌な顔ひとつせずに 夫が 歳暮を 親戚に 贈ってくれるのよ。
RG twitter-pzero:	貴方の**奥さん**優しいなあ。
+ **DCZAR (ours)**:	それは素晴らしい**旦那さん**ね。
対話例 3	
話者 A:	母に具合が悪いことを伝えたら、 母が 病院まで車で送ってくれました。
RG twitter-cloze:	それはありがたいね。
+ **DCZAR (ours)**:	お母さま、優しいですね。

魅力的な応答の特徴は何か？　対話例 3 は，提案モデルの応答がより魅力的[10]であると評価された例の 1 つである．対話例 3 より，魅力的だと評価される応答は，より具体的で首尾一貫した応答であると考えられる．そこで，首尾一貫した応答ほど魅力的な応答であるという仮説を立て，各指標間の相関について分析した．その結果，「文法的流暢性」と「魅力度」における相関係数は 0.223 であり，「首尾一貫性」と「魅力度」における相関係数 0.850 であった．このことから，首尾一貫している応答と魅力的な応答には強い相関があることが示され，内容が一貫した話者と対話を継続したいと感じる傾向にあることが明らかになった．

6)　https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-pages-articles.xml.bz2
7)　https://developer.twitter.com/en/products/twitter-api
8)　どちらの応答が文法にかなっているか，日本語として流暢か (対話履歴の内容は考慮しない)．
9)　どちらの応答が対話履歴の内容を考慮しているか，対話履歴と応答の内容が一貫しているか．
10)　どちらの応答がより魅力的か，どちらの応答をする相手と対話を継続したいと感じるか．