情報源のアノテーションによる外部知識に基づいた応答の魅力度の分析

児玉 貴志 [1]　清丸 寛一 [1]　Yin-Jou Huang[1]　岡久 太郎 [2]　黒橋 禎夫 [1]
[1] 京都大学　[2] 静岡大学
{kodama,kiyomaru,huang,kuro}@nlp.ist.i.kyoto-u.ac.jp
okahisa-taro@inf.shizuoka.ac.jp

概要

人間は何らかの外部知識を参照しながら話す場合でも，自身の知識や意見を適度に織り交ぜながら魅力的に発話を行う．本研究ではそういった人間の振る舞いを，既存の外部知識に基づいた対話コーパスに追加で情報源のアノテーションをすることで分析する．具体的には，発話中の各エンティティに対して，そのエンティティが外部知識に由来する（外部知識由来）か，話者の知識や意見に由来する（話者由来）かをアノテーションする．分析の結果，話者由来情報を含む発話は応答の魅力度が高いことが分かった．また，既存の外部知識に基づく対話応答生成モデルが生成する応答は人間の応答より話者由来情報の割合が少ないことを実験的に確認した．

1　はじめに

対話システムに情報性の高い応答を生成させるために外部知識を活用する研究が増加している [1, 2, 3, 4, 5]．そうした研究では与えられた外部知識を正確に応答に反映させる方法に注目することが多い [6, 7]．

しかし外部知識に基づく対話において，図 1 に示すように，人間は単に外部知識による情報を相手に知らせるだけでなく，話者自身が持っている知識や意見を有効に織り交ぜる [8] ことで，発話を魅力的なものにしている．与えられた外部知識を応答に正確に反映することに特化したモデルが，そのような魅力的な振る舞いをどの程度実現できるかはこれまで定量的に確認されていない．

そこで本研究ではまず，外部知識に基づいた対話コーパスの発話に情報源のアノテーションを追加で施すことで，人間が話者由来情報をどの程度発話に取り込んでいるかを分析する．具体的には，発話中の各エンティティに対して，そのエンティティが外部知識に由来する（外部知識由来）か，話者自身の知識や意見に由来する（話者由来）かをアノテーションする．アノテーションされたデータセットを分析した結果，魅力的な発話は話者由来情報を多く含むことを確認できた．

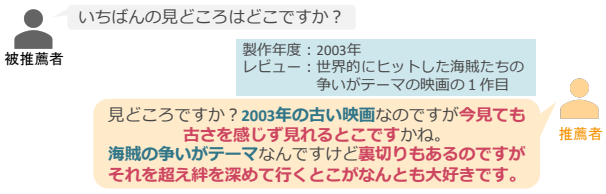

図 1　Japanese Movie Recommendation Dialogue [9] の対話例．推薦者発話中の青字と赤字の箇所がそれぞれ外部知識由来情報と話者由来情報を指す．推薦者発話の上部の四角部分には推薦者が該当発話で使用している外部知識を示す．

次に，パープレキシティを最小化するという一般的な手法で訓練した，BART ベースの応答生成モデルが話者由来情報をどの程度使用できているかを調査した．その結果応答生成モデルは人間ほど頻繁には話者由来情報を使用できていないことが明らかになった．以上の分析と実験結果より，外部知識に基づく応答生成における魅力度を高めるにはパープレキシティを最小化するだけでは不十分であり，学習フレームワークに改善の余地があることが示唆された．

2　情報源のアノテーション

本節では情報源ラベルのアノテーション方法とその結果について説明する．

2.1　アノテーション方法

外部知識に基づいた対話コーパスである Japanese Movie Recommendation Dialogue (JMRD) [9] に追加で情報源のアノテーションを行う．JMRD は映画推薦

についての日本語対話コーパスで，推薦者は発話を行う際に，参照した映画情報にアノテーションを行っている．この手続きにより推薦者の全ての発話に，外部知識として映画情報が紐付けられている．各知識は知識タイプ（タイトル，製作年度，監督，キャスト，ジャンル，レビュー，あらすじ，知識なしの 8 つ）と対応する知識コンテンツから構成されている．表 7 に JMRD の対話例を，表 8 に JMRD で使用されている外部知識の例を示す．

本研究では推薦者の発話中の全ての体言，用言，形容詞をアノテーションの対象とし，これらをエンティティと呼ぶ．エンティティの抽出には，形態素解析器 Juman++ [10] を使用する．抽出の際にはエンティティの意味を捉えやすくするため，エンティティを修飾している形容詞や副詞も含めて抽出する．アノテータは抽出されたアノテーション対象を以下に示す情報源のタイプのいずれかに分類する．

外部知識由来： 対象のエンティティがその発話で使用されている外部知識に基づいている．
話者由来： 対象のエンティティが推薦者が推薦映画について元々持っている知識や意見に基づいている．
その他： 上記 2 つに該当しない．

以下に分類の実例を示す．

(1) 発話：アクションシーン (外部知識由来) は
見ごたえ (話者由来) 抜群です
使用されている知識：{ジャンル, アクション}

アノテーションはプロのアノテータに依頼した．1 人のアノテータが 1 つの対話を担当し，アノテーション後に別のアノテータが内容を確認した．アノテーションにかかった費用は約 140 万円だった．

2.2 アノテーション結果

表 1 にアノテーションの統計を示す．推薦者は外部知識を参照しながら発話をしているため，その性質上外部知識由来のエンティティが多いものの，話者由来のエンティティも約 6 万個存在した．この結果は，人間は外部知識を伝えるための対話であっても，自身の知識や経験，意見を発話に取り入れていることを示している．

表 1 情報源のアノテーションの統計

	学習	開発	テスト	合計
対話	4,575	200	300	5,075
推薦者発話	51,080	2,244	3,347	56,671
エンティティ	235,771	10,320	15,734	261,825
外部知識由来	166,958	7,223	10,476	184,657
話者由来	51,170	2,303	4,095	57,568
その他	17,643	794	1,163	19,600

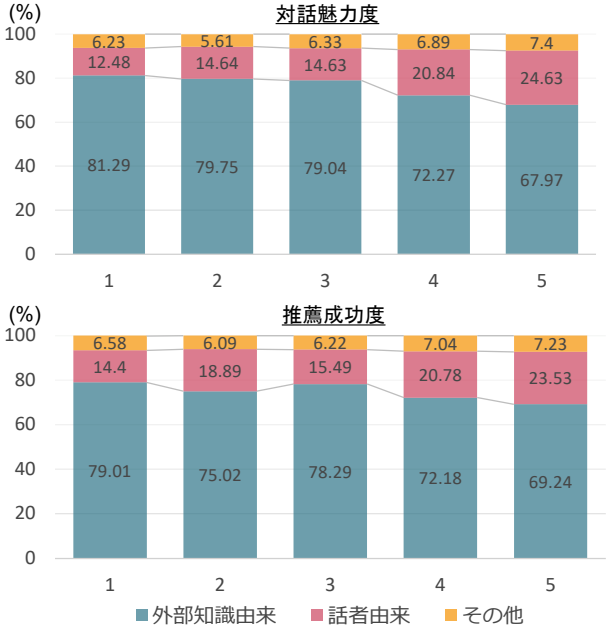

図 2 アンケート結果と各情報源ラベルとの関係

3 人間の応答の分析

本節では人間の発話の魅力度と話者由来情報の関係を対話レベルと発話レベルに分けて分析する．

3.1 対話レベルの分析

JMRD の一部の対話（4,317 対話）には対話終了後に取った 5 段階評価のアンケート（5 が最良）が付随している．そのアンケートの中から，被推薦者側の，対話魅力度（質問文：「対話を楽しめたか」）と，推薦成功度（質問文：「推薦された映画を見たくなったか」）の 2 つのアンケート結果と各情報源ラベルの割合との関係について分析する．

図 2 にアンケートのスコア別に各情報源ラベルの割合を示す．この図より対話魅力度，推薦成功度ともに，評価の低い対話では話者由来情報が少なく（または外部知識由来情報が多く），評価の高い対話では話者由来情報が多い（または外部知識由来情報が少ない）傾向が見て取れる．JMRD ではクラウド

表 2 人間の応答に対する発話魅力度[1]．括弧内の数値は応答数を指す．

	発話魅力度
話者由来情報あり (230)	3.31
話者由来情報なし (266)	3.07
合計 (496)	3.18

ワーカーに一定量の外部知識を提示し，その知識を使いながら対話をするように教示している．ただ，対話相手からの評価が高い推薦者は，与えられた外部知識だけではなく，話者由来の知識をある程度（統計的には約 25% 程度）混ぜながら発話をしていることが分かった．

3.2 発話レベルの分析

JMRD には発話単位では応答の魅力度が付随していないため，新たにクラウドソーシングで応答の魅力度（発話魅力度）について評価を行った．まずテストデータの中から 500 個の対話文脈（4 発話）とその文脈に対する応答を抽出し，クラウドワーカーにその応答の発話魅力度（質問文：「この応答をした人と話してみたい」）を 5 段階評価（5 が最良）で評価してもらう．各応答につき 3 名のクラウドワーカーに評価してもらい，スコアはその平均とする．

表 2 に結果を示す．話者由来情報を含む応答の平均スコアが 3.31 であるのに対し，含まない応答の平均スコアは 3.07 であった．この結果より，話者由来情報を含む応答の方がより魅力的であることが発話レベルでも示された．

4 システム応答の分析

本節では一般的な手法で訓練した応答生成モデルが生成した応答中の情報源ラベルの分布を調査する．まず対話文脈と外部知識から応答を生成する応答生成器（4.1 節）を訓練する．次に応答と外部知識を入力，情報源ラベルの BIO タグを出力とする情報源分類器（4.2 節）を訓練する．最後に応答生成器が生成した応答（システム応答と呼ぶ）に対して，訓練した情報源分類器によって情報源ラベルを推測して付与し，その分布を確認する．

4.1 応答生成器

応答生成器は日本語 Wikipedia を使用して事前学習された BART$_{large}$ [11] を fine-tuning して作成する[2]．モデルへの入力は以下のように定める．

$$[CLS]u_{t-4}[SEP]u_{t-3}[SEP]u_{t-2}[SEP] u_{t-1}[SEP][CLS_K]kt^1[SEP]kc^1[SEP]... [CLS_K]kt^M[SEP]kc^M[SEP], \quad (1)$$

t は対話ターン，u_t は t 番目の発話，kt^i と kc^i $(1 <= i <= M)$ は応答 u_t に紐付けられている知識タイプと知識コンテンツである（M は応答に紐付けられている知識の最大数）．$[CLS_K]$ は特殊トークンである．実運用時は知識選択も行う必要があるが，本研究では知識が応答にどのように反映されているかに注目するため，知識選択タスクは行わず，モデルには常に正解の知識を入力する．モデルは応答 u_t のパープレキシティ最小化を目的関数として学習する．

システム応答の評価には SacreBLEU [12] を使用した．正解の知識を与えているため，BLEU-1/2/3/4 は 81.1/73.5/71.0/69.9 と非常に高いスコアとなった．3.2 節で行った，発話魅力度の評価をシステム応答に対しても同様に行った．平均スコアは 3.09 であり，人間と比較するとやや低いスコアとなった．

表 3 情報源分類器による系列ラベリング結果

	Prec	Rec	F1
外部知識由来	94.92	95.61	95.27
話者由来	80.88	84.39	82.60
その他	82.93	64.15	72.34
マイクロ平均	90.52	90.48	90.50

4.2 情報源分類器

情報源分類器は日本語 Wikipedia と CC-100 を使って事前学習された RoBERTa$_{large}$ [13] を fine-tuning して作成する[3]．情報源分類器は，情報源の BIO ラベルを推定する系列ラベリングタスクを解く．モデルへの入力は以下のように定める．

$$[CLS]u_t[SEP][CLS_K]kt^1[SEP]kc^1[SEP]... [CLS_K]kt^M[SEP]kc^M[SEP] \quad (2)$$

表 3 に各ラベルの適合率 (Prec)，再現率 (Rec)，F1 スコア (F1) とそれぞれのマイクロ平均スコアを示す．マイクロ平均 F1 スコアは 90.50 であり，後段の分析に使用するには十分な精度である．

1) 情報源ラベルのない 4 発話についてはこの結果から除外している．

2) https://nlp.ist.i.kyoto-u.ac.jp/?BART%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB

3) https://huggingface.co/nlp-waseda/roberta-large-japanese-seq512

表 4　人間とシステムの応答例．外部知識の行では，波括弧内の左側が知識タイプ，右側が知識コンテンツである．

		発話魅力度
文脈	... 推薦者: 2015 年のアニメです． 被推薦者: なるほど〜〜	
外部知識	{監督, 京極尚彦}, {キャスト, 新田恵海}, {キャスト, 南條愛乃}	
応答	人間: 監督は京極尚彦，声は新田恵海，南條愛乃です．この二人，歌手もやってます． システム: 監督は京極尚彦，声は新田恵海，南條愛乃です．	4.00 2.33

表 5　人間とシステム応答の情報源ラベルの分布．gold はアノテータによって付与されたラベルを，pred は情報源分類器で推定したラベルを指す．

分布 (%)	Human (gold)	Human (pred)	System (pred)
外部知識	66.22	66.75	85.48
話者	26.33	27.49	10.66
その他	7.45	5.77	3.86

表 6　知識タイプ別の話者由来情報の割合

割合 (%)	Human (gold)	Human (pred)	System (pred)
タイトル	30.21	34.12	27.09
製作年度	16.41	22.31	6.56
監督	13.94	11.96	4.50
キャスト	36.11	45.34	23.45
ジャンル	10.47	15.14	5.49
レビュー	27.72	31.42	6.32
あらすじ	13.98	13.68	2.32
知識なし	57.49	63.08	55.99

4.3　推定ラベルに対する分析

4.2 節で訓練した情報源分類器を用いて，システム応答の情報源ラベルを推定する．表 5 に人間とシステム応答の情報源ラベルの分布を示す．人間の応答にはアノテーション（2 節）ですでに正解ラベルが付与されている（**Human (gold)**）が，公平な比較のため人間の応答についても情報源分類器で推定ラベルを付与する（**Human (pred)**）．**Human (gold)** と **Human (pred)** は似た分布になっており，分類器の精度が十分高いことを示している．

推定したラベルが付与されたシステム応答である **System (pred)** は，**Human (pred)** と比較すると，外部知識由来の割合が大きく増加（66.75%→85.48%）し，話者由来の割合が大きく減少（27.49%→10.66%）した．この結果より，一般的な手法で訓練した応答生成モデルは人間より話者由来情報の使用量が減少してしまうことが明らかになった．

表 4 に，人間とシステムの応答例を発話魅力度とともに示す．システムは外部知識を適切に反映しているものの，人間の応答中の「この二人，歌手もやってます．」のような，話者由来情報を追加するといった振る舞いはできていない．

さらなる分析として，表 6 に知識タイプ別の話者由来情報の割合を示す．人間の発話と比較すると，システム発話の話者由来情報の量はすべての知識タイプで減少した．特にレビュー（31.42%→6.32%）やあらすじ（13.68%→2.32%）では大きな減少が見られた．レビューやあらすじは誰かの知識や意見を記述した外部知識であり，システムはこうした外部知識を参照することで，他に話者由来情報を追加する必要がなかったためであると考えられる．

話者由来情報が応答の魅力度を向上させるという分析と合わせると，現在の主流のモデルは話者由来情報を効果的に取り込むことができないため，その魅力度が低下してしまっている可能性がある．こうした能力は応答生成のパープレキシティを最小化するようにモデルを最適化するだけではほとんど学習されないため，新たな学習フレームワークが必要であると考えられる．

5　おわりに

本研究では，外部知識に基づいた対話における，人間およびシステム応答に含まれる話者由来情報について分析した．分析の結果，外部知識だけでなく話者由来情報を用いることで，より魅力的な応答が得られることが分かった．また標準的な方法で学習した応答生成モデルは，人間よりも話者由来情報の生成量が少ないことを定量的に確認した．

対話に出現する話者由来情報は多種多様であり，単純に応答のパープレキシティを最小化するだけでは話者由来情報をうまく活用することは困難である．今回施したアノテーションは公開予定であり，このアノテーション付きコーパスがこの問題に取り組む上での良い出発点となることを期待する．

謝辞

この研究は国立情報学研究所 (NII) CRIS と LINE 株式会社とが推進する NII CRIS 共同研究の助成，JST，CREST，JPMJCR20D の支援，及び JSPS 科研費 JP22J15317 の助成のもとで行われた．

参考文献

[1] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, April 2018.

[2] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 708–713, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[3] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. Towards exploiting background knowledge for building conversation systems. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2322–2332, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In International Conference on Learning Representations, 2019.

[5] 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 (EMNLP), pp. 3377–3390, Online, November 2020. Association for Computational Linguistics.

[6] Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. Sequential latent knowledge selection for knowledge-grounded dialogue. In International Conference on Learning Representations, 2020.

[7] Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. Augmenting knowledge-grounded conversations with sequential knowledge transition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5621–5630, Online, June 2021. Association for Computational Linguistics.

[8] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 704–718, Online, August 2021. Association for Computational Linguistics.

[9] Takashi Kodama, Ribeka Tanaka, and Sadao Kurohashi. Construction of hierarchical structured knowledge-based recommendation dialogue dataset and dialogue system. In Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pp. 83–92, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Design and structure of the Juman++ morphological analyzer toolkit. Journal of Natural Language Processing, Vol. 27, No. 1, pp. 89–132, 2020.

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

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

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. Vol. abs/1907.11692, , 2019.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

表 7 JMRD の対話例．話者列の R と S はそれぞれ推薦者と被推薦者を指し，添字の番号は対話のターン数を示す．「知識なし」は推薦者が外部知識を使用しなかったことを示す．

話者	対話	知識タイプ	知識コンテンツ
R_1	こんにちは	知識なし	-
S_2	こんにちは．よろしくお願いします！		
R_3	アベンジャーズ/エンドゲームは知っていますか？	タイトル	アベンジャーズ/エンドゲーム
S_4	タイトルを聞いたことがある程度です・・・		
R_5	この映画は 2019 年に公開された映画です	製作年度	2019
S_6	なるほど，アメリカの映画ですか？		
R_7	アメリカのアクション映画です	ジャンル	アクション
S_8	見どころはどのようなところでしょうか？		
R_9	悪役のサノスという星人がいるのですが，大集結してサノスに立ち向かうところがみどころです	レビュー	大集結してサノスに立ち向かうところ
S_{10}	なるほど！ 宇宙で戦いが繰り広げられるストーリーなのですか？		
R_{11}	いや，舞台は地球です	知識なし	-
S_{12}	となると，地球に悪役が攻めてくるのですね・・・.		
R_{13}	そうですね，結構怖い場面もあります	知識なし	-
S_{14}	怖いのですか・・・ 私はホラー系は苦手ですが，アクション系は好きです． 私のような場合でも楽しんで見られるでしょうか？		
R_{15}	ホラーのような怖さはないので，楽しんで見られると思います	知識なし	-
S_{16}	なるほど！ サノスとヒーローとの闘い，ワクワクしそうですね！		
R_{17}	ぜひ見てください！	知識なし	-
S_{18}	はい！ 近々レンタルビデオ店に行く機会があるので，アベンジャーズ/エンドゲームをレンタルしてみたいと思います！		
R_{19}	ありがとうございました	知識なし	-
S_{20}	こちらこそ，貴重な情報ありがとうございました！		

表 8 JMRD で使用されている外部知識の一例．監督とキャストには名前と説明がそれぞれ存在する．

知識タイプ		知識コンテンツ
タイトル		アベンジャーズ/エンドゲーム
製作年度		2019
監督	名前	アンソニー・ルッソとジョー・ルッソ
	説明	アメリカ合衆国のテレビ及び映画の監督，プロデューサー，脚本家，俳優，編集技師.
キャスト	キャスト $_1$ 名前	ロバート・ダウニー・Jr
	キャスト $_1$ 説明	アメリカ合衆国の俳優・声優・ミュージシャン・プロデューサー
	キャスト $_2$ 名前	クリス・エヴァンス
	キャスト $_2$ 説明	アメリカ合衆国の俳優．マサチューセッツ州サドベリー出身.
ジャンル		アクション，アドベンチャー
レビュー		「インフィニティーサーガの最終章でサノスとの決戦が描かれている注目作品」など 5 文
あらすじ		「2018 年，タイタン星人サノスによるデシメーション（インフィニティ・ストーンの力を使った大量殺戮）で全宇宙の生命の半分が消し去られてから 3 週間.」など 10 文

表 9 応答生成器と情報源分類器のハイパーパラメータ

パラメータ名	応答生成器	情報源分類器
エポック数	50	20
Early stopping の patience	3	
バッチサイズ	512	64
最大入力トークン長（対話文脈/外部知識/応答）	256 / 256 / 128	- / 384 / 128
オプティマイザ	AdamW [14]	
学習率	5e-5	
ウォームアップのステップ数	1000	
勾配クリッピングの閾値	0.5	
ビームサイズ	3	-