知識グラフの対話システムへの記憶化：学習アプローチの探究

薛 強 [1] 滝口 哲也 [1] 有木 康雄 [1]
[1] 神戸大学システム情報学研究科
xueqiang@stu.kobe-u.ac.jp takigu,ariki@kobe-u.ac.jp

概要

近年，対話システムの進化に伴い，生成される回答の質の向上が重要な課題となっている．本研究では，知識グラフを対話システムに記憶させることにより，回答の質を向上させると同時に，回答中の幻想問題を軽減する方法を探究する．具体的には，知識グラフを記憶するための 3 種類の学習タスクを提案し，その有効性を検証する．実験結果から，一般的な知識検索対話システムと比較して回答の生成品質は向上しなかったものの，新たな可能性を探る一歩となった．本研究は，対話システムの質的向上と知識グラフの活用における新たな方向性を示唆している．

1 はじめに

オープンドメイン対話システムは，人工知能と自然言語処理の分野で重要な進展を遂げてきたが，依然として重大な課題に直面している．具体的には，これらのシステムにはしばしば一貫性のない回答や，事実上の誤りを含む回答を生成するという問題がある．この問題の原因は，文脈の深い理解不足や，特定の知識に基づく正確な回答の生成が困難であることに由来する．

このような背景のもと，知識ベース対話システムが提案され，注目を集めている．これらのシステムは，特定の知識ベースに基づいて情報を提供することで，回答の質と正確性を向上させることができる．たとえば，Young ら [1] や Huang ら [2] の研究では，知識ベースを活用することで，対話システムの回答精度が向上することが示されている．このアプローチは，情報量を増加させると共に，対話の一貫性を保つための重要な手段ともなっている..

一方で，知識検索対話システムは，Dinan ら [3] や Ghazvininejad ら [4] などの研究で示されるように，広範な情報源からのデータ検索に基づいて回答を生成する．これにより，様々なトピックに対応する能力が向上するが，検索エラーや非効率性，関連性の低下などの問題も伴う．

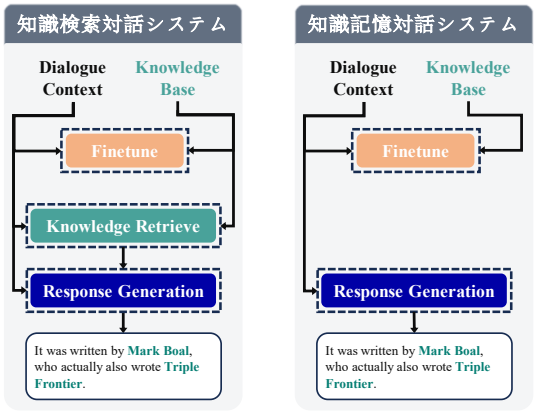

図 1 二種類の知識ベース対話システム

これらの課題に対処するため，言語モデルのパラメータに直接知識を組み込む知識記憶対話システムの研究が進められている．Sun ら [5] の研究は，この新しいアプローチが知識検索プロセスを必要とせずに，高品質な対話を生成できることを示している．このようなシステムは，対話の自然さと応答性を大幅に向上させることが期待できる．

特に，知識グラフに基づく知識ベース対話システムは，構造化された知識形式を活用することで，より洗練された情報提供を可能にする．知識グラフを対話システムに記憶させることは，回答の正確性と関連性を大幅に向上させることができ，ユーザー体験の質を高めることが期待される．本研究では，この分野での新たな進展として，対話システムが知識グラフを効果的に記憶するための 3 つの異なる学習方法を提案し，その実用性と効果を検証する．

本稿では，まず知識検索対話システムと知識記憶対話システムについて述べる．次に提案する対話システムについて述べる．最後に提案する対話システムの実験と評価について報告する．

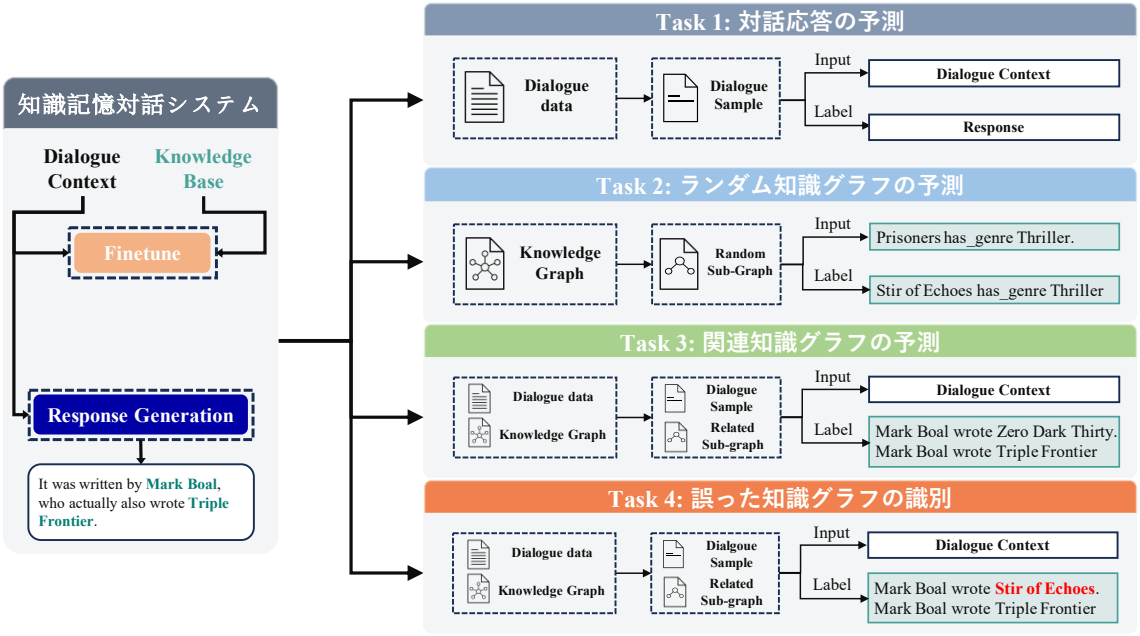

図 2　対話応答生成のための予測タスクと，提案された三つの学習タスクにおけるデータ処理の流れ．

2　関連研究

本章では，二種類の知識ベース対話システムについて述べる．

2.1　知識検索対話システム

オープンドメインの知識ベース対話 (KGD) において，人々は知識統合により意義ある対話を行っている [6, 7]．知識統合を実現するための KB ベース手法が数多く研究され [8, 6, 9]，知識選択 [10, 11] や応答生成 [12, 13] に焦点が当てられている．しかし，これらの手法は検索エラーや非効率性 [14]，多粒度知識の統合問題 [15] などの課題に直面している．

2.2　知識記憶対話システム

言語モデル (LM) は，大量の知識をパラメータに記憶し，会話において有益な応答を生成する能力を有している [16]．LM が知識ベースとして機能することは Petroni らによって示され [17]，これは質問応答タスクなどにおいても有効であることが明らかになっている [18]．その一方で，LM ベースの手法は誤情報，いわゆる「幻覚」の問題を引き起こす可能性があり，これは対話システムの信頼性に影響を及ぼす．MixCL[5] は，主にテキストベースの知識データにおける幻覚を低減する効果があるが，構造化された知識グラフデータに対しては十分な効果を発揮できない．本研究では，知識グラフデータを用いた知識記憶対話システムの性能向上を目指し，幻覚問題を軽減するための異なる学習手法を提案する．

3　提案学習アプローチ

本章では，知識グラフを知識記憶対話システムに覚えさせるために，提案する対話システムが用いる四つの学習アプローチについて述べる．

3.1　Task 1：対話応答の予測

対話履歴を C，目標とする応答を $(x_1, \ldots, x_L)$ とした場合，モデルの学習は，以下の負の対数尤度を最小化することにより行われる．

$$L_{\mathrm{LM}} = -\sum_{i=1}^{L} \log P(x_i \mid C, x_1, \ldots, x_{i-1})$$

学習された生成言語モデルを利用し，対話システムは適切な応答を生成する能力を獲得する．

3.2　タスク 2：ランダム知識グラフの予測

このタスクでは、知識グラフ G から抽出されたサブグラフ G_{sub} を注目し、このサブグラフを知識トリプルのテキスト形式の集合として表現する。各知識トリプル g_i は (w_{i1}, w_{i2}, w_{i3}) として表され、ここで w_{i1}、w_{i2}、w_{i3} はそれぞれトリプルの要素（エンティティや関係）を表す単語である。サブグラフの長さを L とし、モデルの入力はサブグラフ G_{sub} の先頭からランダムに選択された位置 $r-1$ までの知識トリプルの集合 $(g_1, \ldots, g_{r-1})$ となる。目標はこの入力に

基づいて、続く知識トリプルの集合 $(g_r, \ldots, g_L)$ を予測する。ここで、r は 1 と L の間でランダムに選ばれるインデックスである。モデルの学習は，以下の負の対数尤度を最小化することにより行われる．

$$L_{\text{RKG}} = -\sum_{i=r}^{L} \log P(g_i \mid g_1, g_2, ..., g_{i-1})$$

3.3 タスク 3：関連知識グラフの予測

本タスクでは，対話データと関連する知識グラフを組み合わせることにより，与えられた対話コンテキストに基づいて最も関連性の高いサブグラフを予測することを目的としている．対話履歴を C，関連する知識グラフのサブグラフ G_{sub} を知識トリプルのテキスト形式の集合 $(g_1, \ldots, g_{r-1})$ として表現する．タスクの入力として対話履歴 C が与えられ，目標として知識グラフから関連するサブグラフ G_{sub} を予測する．モデルの学習は，以下の負の対数尤度を最小化することにより行われる．

$$L_{\text{CKG}} = -\sum_{i=1}^{L} \log P(g_i \mid C, g_1, g_2, ..., g_{i-1})$$

3.4 タスク 4：誤った知識グラフの識別

このタスクでは，Sun による研究 [5] を参照し，対話システムにおける「幻覚」の問題を緩和するために，対照学習手法を採用している．タスク 3 で生成された関連知識グラフ G_{sub} から，エンティティや関係をランダムに別のエンティティや関係に置換し，その結果得られる誤った知識を負の例として使用する．モデルは，誤った知識 (負の例) の対数尤度を最小化し，同時に正しい知識 (正の例) の対数尤度を最大化するように学習される．

このプロセスを数式化すると，次のようになる．

$$L_{\text{CL}} = -\sum_{i=1}^{L} \left[\log P(g_i \mid C, G_{\text{pos}}) - \log P(g_i \mid C, G_{\text{neg}})\right]$$

ここで，C は対話コンテキスト，G_{pos} は正しい知識の集合，G_{neg} は誤った知識の集合を表している．このようにして，モデルは誤った情報を識別し，正しい情報を強化することを学習する．

全体の損失関数は，各タスクに対する損失の加重和である．

$$L_{\text{Total}} = \alpha L_{\text{LM}} + \beta L_{\text{CKG}} + \gamma L_{\text{RKG}} + \delta L_{\text{CL}}$$

ここで，α，β，γ，δ は各損失関数の重要度を調整する超パラメータである．

4 実験と評価

本章では，以下 2 つの角度から提案手法の性能を考察するための実験について述べる：

- 異なる対話システムの性能比較: 4.2 節
- 提案する異なるタスクの性能比較: 4.3 節

4.1 実験設定

実験データセットは OpenDialKG [19] を用いる．OpenDialKG は，本と映画についての推薦対話が含まれている雑談対話データセットである．話者は常に構造化知識中のエンティティを含む発話を行い，関連知識に基づいた推薦対話が行われている．実験で用いたハイパーパラメータの設定を表 1 に示す．

実験評価では，正解性と多様性の二つの角度から応答文の質を評価する．応答正解性の評価指標としては，応答と正解文の F1 スコア（F1），応答文と正解文の類似度を表す BLEU-n (B2,B4) [20]， Meteor [21] を用いる．応答多様性の評価指標として，応答文に含まれる n-gram の種類数を表す DIST-n（D2, D3）[22] を用いる．応答に含まれる知識の正解性の評価指標としては，応答知識と正解知識の F1 スコア（KF1）と，応答エンティティと正解エンティティに対する F1 スコア（EF1）を用いる．

4.2 対話システムの比較実験

本実験では以下三つの対話システムを比較対象とする：

- **知識なし対話システム**：知識データを使用せず，対話の自然さと流暢さを重視したシステム．
- **知識検索対話システム**：目標知識データに基づいて学習と推論を行い，知識を取り入れたシステム．
- **知識記憶対話システム**：本研究で提案するアプローチであり，知識グラフのデータを直接言語モデルのパラメータに組み込むことで，推論時の知識検索を必要としない．

表 1 ハイパーパラメータ

モデル	BART
最適化アルゴリズム	AdamW
最大対話履歴長	256 tokens
Epochs	10
Beam Size	12
Learning Rate	6.0e-5
α,β,γ,δ	0.4,0.2,0.2,0.3

表 2　各モデルによって生成された応答文の評価結果.

System	Model	応答正解性				応答多様性		知識正解性	
		F1 ↑	B2 ↑	B4 ↑	ME ↑	D2 ↑	D3 ↑	KF1 ↑	EF1 ↑
知識なし	BART	23.11	10.77	3.26	21.52	15.42	25.69	6.73	12.76
知識検索	KG-BART	**27.82**	**13.83**	**5.29**	**25.14**	18.18	29.33	**17.31**	**28.42**
知識記憶	Ours	23.24	10.91	3.51	21.63	**18.44**	**31.03**	7.50	15.42
知識記憶	w/o L_{RKG}	23.00	10.83	3.44	21.31	16.91	27.60	8.28	14.67
	w/o L_{CKG}	23.24	11.05	3.36	21.97	14.94	24.92	6.58	12.66
	w/o L_{CL}	23.47	10.95	3.26	21.65	16.24	26.61	7.97	14.82

表 3　各モデルによる対話例．赤字は応答中にある問題単語.

生成例 1	Knowledge	None.
	User A	Can you recommend work by Jason Statham?
	User B	I can recommend you the movie Wild Card or The Transporter
	User A	Sure, I would like to know more about Wild Card please.
	User B	Wild Card has genre Crime/Action and was directed by Simon West
	User A	Okay cool, do you know when it was released?
BART		Sorry I do not have that information.
KG-BART		Sorry I don't have that information.
Ours		It was released in 2012.
生成例 2	Knowledge	Becca Fitzpatrick written_by Silence
	User A	Do you know Becca Fitzpatrick?
BART		Yes, she is a great author. Have you read any of her books?
KG-BART		Yes, she wrote Silence. Have you read it?
Ours		Becca Fitzpatrick wrote Silence and Hush, Hush.

各対話システムの応答文生成評価結果を表 2 に示す．提案システムは，特に応答多様性の指標（D2, D3）で優れた性能を発揮しており，これはシステムが生成する応答の幅広さと創造性を反映している．知識正解性（KF1, EF1）においては，現段階では知識検索対話システムに劣るものの，構造化された知識ベースとの連携強化など，今後の改良が期待できる．

表 3 の対話例 1 において，知識がない状態では，知識検索対話システムは「Sorry I don't have that information.」と回答し，利用可能な情報がない場合の限界を示している．一方，提案システムは検索に依存しないが，不正確な情報を生成することがあり，「2012」という幻覚に陥っている．この問題は，提案システムが知識を内部化しているものの，正確性を保証するメカニズムがまだ不十分であることを示しており，今後の改善が必要であることを示唆している．生成例 2 では、知識検索対話システムは提供された知識を利用して適切な応答を生成し、提案システムは学習で得た知識を活用して、検索を必要とせずに関連する回答を生成している。

4.3　アブレーション実験

本実験では，3 章で述べた異なるタスクの性能を比較した．表 2 下半部分に，異なるタスクの構成で学習されたシステムによって生成された応答文の評価結果を示す．L_{RKG} を除外した場合、応答正解性と知識正解性の両方で性能が低下しており、この損失がシステムの全体的な性能向上に寄与していることが示されている。また、L_{CKG} を除外すると、応答多様性が低下し、知識正解性にも悪影響を与えていることが観察される。一方で、L_{CL} を除外した場合は、応答正解性にはあまり影響がないものの、応答多様性と知識正解性には若干の低下が見られる。

5　おわりに

本研究では，知識検索に依存せずに動作するオープンドメイン対話システムの知識記憶アプローチを探求した．提案システムは応答の多様性と流暢さを向上させるものの，実験結果は幻覚の問題を浮き彫りにした．これは知識統合の過程での課題であり，今後の改善の必要性を示している．

謝辞

本研究の一部は，JSPS 科研費 JP21H00906 の支援を受けたものである．

参考文献

[1] Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. Augmenting end-to-end dialog systems with commonsense knowledge. **Proceedings of AAAI**, 2018.

[2] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. **ACM Transactions on Information Systems (TOIS)**, Vol. 38, No. 3, p. 1–32, 2020.

[3] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Vlad Serban, Ryan Lowe, et al. The second conversational intelligence challenge (convai2). In **Proceedings of NeurIPS Conversational AI Workshop**, 2019.

[4] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. A knowledge-grounded neural conversation model, 2018.

[5] Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, M. de Rijke, and Zhaochun Ren. Contrastive learning reduces hallucination in conversations. In **AAAI Conference on Artificial Intelligence**, pp. 13618–13626, 2023.

[6] Tom Young, E. Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialog systems with commonsense knowledge. In **AAAI**, 2018.

[7] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. **ACM TOIS**, Vol. 38, pp. 1 – 32, 2020.

[8] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In **ACL**, pp. 74–81, 2018.

[9] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In **ICLR**, 2019.

[10] Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and M. de Rijke. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In **SIGIR**, pp. 1151–1160, 2020.

[11] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In **Findings of EMNLP**, pp. 3784–3803, 2021.

[12] Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. Knowledge-grounded dialogue generation with pre-trained language models. In **EMNLP**, pp. 3377–3390, 2020.

[13] Chujie Zheng and Minlie Huang. Exploring prompt-based few-shot learning for grounded dialog generation, 2021.

[14] Yan Xu, Etsuko Ishii, Zihan Liu, Genta Indra Winata, Dan Su, Andrea Madotto, and Pascale Fung. Retrieval-free knowledge-grounded dialogue response generation with adapters. In **ACL — Workshop on DialDoc**, pp. 93–107, 2022.

[15] Zhiyong Wu, Wei Bi, Xiang Li, Lingpeng Kong, and Benjamin C.M. Kao. Lexical knowledge internalization for neural dialog generation. In **ACL**, pp. 7945–7958, 2022.

[16] Yufan Zhao, Wei Wu, and Can Xu. Are pre-trained language models knowledgeable to ground open domain dialogues?, 2020.

[17] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? In **EMNLP**, pp. 2463–2473, 2019.

[18] Adam Roberts, Colin Raffel, and Noam M. Shazeer. How much knowledge can you pack into the parameters of a language model? In **EMNLP**, pp. 5418–5426, 2020.

[19] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 845–854, 2019.

[20] 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.

[21] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. 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. **arXiv preprint arXiv:1510.03055**, 2015.