JMultiWOZ に対する対話状態アノテーションの付与と対話システムの実装評価

大橋厚元 * 平井龍 * 飯塚慎也 東中竜一郎
名古屋大学大学院情報学研究科
{ohashi.atsumoto.c0, hirai.ryu.k6, iizuka.shinya.a8}@s.mail.nagoya-u.ac.jp
higashinaka@i.nagoya-u.ac.jp

概要

日本語のマルチドメインタスク指向型対話データセットとして JMultiWOZ が構築されている．しかし，対話状態のアノテーションが付与されていないため，対話モデルの構築には利用できない．本研究では新たに，対話状態のアノテーションを JMultiWOZ に対して追加することで，対話状態追跡と応答生成を遂行できる対話モデルの構築を目指す．さらに，本データセットを用いて実装された対話モデルの評価実験を実施し，英語圏の標準的なデータセットである MultiWOZ2.2 と同難易度のベンチマークを，JMultiWOZ が提供できることを示す[1)]．

1 はじめに

タスク指向型対話システムの研究では，ニューラルモデルをベースとした手法の活発な導入 [1, 2] によって，対話能力の進化が著しい [3, 4]．このような進化には，ニューラルモデルを学習するための対話データセットが必須である．英語圏では，現在までに構築された数多くの対話データセットが，深層学習を用いた対話モデルの発展に貢献してきた [5, 6]．

MultiWOZ [7] は，現在英語圏において最も用いられているタスク指向型対話データセットの一つであり，旅行に関する 7 つのドメインにまたがる対話を 1 万件以上収録している大規模コーパスである．MultiWOZ を先駆けとして，より複雑で大規模なタスク指向型対話データセットも多く構築されており [8, 9, 10]，近年の対話モデルの発展を牽引している [11, 12, 13]．また，中国語圏においても，複数のマルチドメインタスク指向型対話データセットが構築 [14, 15, 16] され，中国語タスク指向型対話システムの発展に寄与している．

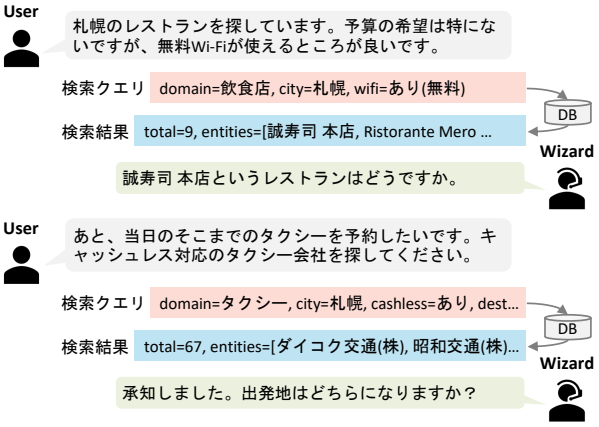


図 1 JMultiWOZ に含まれる，2 ドメイン（飲食店とタクシー）にまたがる対話の例．灰色と緑色のテキストはそれぞれ旅行者役 (user) と情報提供者役 (wizard) の発話を示している．また，赤色と青色のボックスは，wizard が入力したデータベース検索クエリとその結果を示している．

先行研究において我々は，日本語によるタスク指向型対話システムの研究開発の促進を目指し，日本語初のマルチドメインタスク指向型対話データセット JMultiWOZ を構築した [17]（対話例を図 1 に示す）．そして，検索クエリを推定するモデルを学習，評価することで，本データセットの有用性を示した．しかし，現状の JMultiWOZ には対話状態ラベルのアノテーションが付与されておらず，タスク指向型対話における主要なタスク，すなわち，対話状態追跡 (dialogue state tracking; DST) と応答生成 (response generation; RG) のモデリングが行えない，という制限があった．これは，End-to-End な対話システムを構築できないことを意味する．

本研究では，日本語における DST および RG のベンチマークを提供するため，先行研究における JMultiWOZ に対して新たに対話状態アノテーションを追加する．そして，MultiWOZ における state-of-the-art (SOTA) モデル [18] と，最新の LLM

* Equal contribution
1） JMultiWOZ と実験ソースコードは，https://github.com/nu-dialogue/jmultiwoz で公開されている．

表 1 Wizard が入力したデータベース検索クエリと，追加のアノテーションによって得られる対話状態ラベルの例．**太字**は，検索クエリと対話状態の差分を示している．

文脈	User:	札幌のホテルを探していて、Wi-Fi が無料で使えると助かります。
	Wizard:	では、JR INN 札幌はどうでしょうか？
	User:	よさそうですね，いくらくらいでしょうか？
検索クエリ	city: 札幌, wifi: 有り（無料）	
対話状態	city: 札幌, wifi: 有り（無料）, **name: JR INN 札幌**	

ベースのモデル [19] を用いて，JMultiWOZ の DST および RG を学習し，本データセットが MultiWOZ と同難易度のベンチマークを提供できることを実証する．さらに，これら対話モデルの人間評価実験を実施し，最新の LLM であっても日本語におけるタスク指向型対話の能力には課題があることを示す．

2 JMultiWOZ の概要

JMultiWOZ は，日本の 9 都市（札幌，仙台，東京，横浜，名古屋，京都，大阪，福岡，那覇）のいずれかへの旅行者が，観光情報を収集しながら旅行を計画する対話を，合計で 4,246 対話収録したデータセットである [7]．対話のドメインは，観光名所, 宿泊施設, 飲食店, 買い物施設, タクシー, 天気の 6 つにわたる．各対話は，Wizard-of-OZ 法 [20] に基づき，旅行者役 (user) と情報提供者役 (wizard) のクラウドワーカによって実施された．対話の全てのターンには，user の要求に基づいて wizard が検索したデータベースの検索クエリとその検索結果が記録されている．

3 対話状態アノテーション

対話状態とは，各ターン時点までに判明している，ユーザが求めるエンティティ（具体的な観光名所や飲食店のこと）の条件を記録した情報であり，スロットと値のペアの集合で定義される．基本的には，対話収集時に wizard が入力したデータベース検索クエリが対話状態の一部として利用できる．しかし，ユーザが伝達したい要求をすべてカバーするためには，検索クエリには反映されなかった情報も対話状態に含める必要がある．

例えば，表 1 に示されるような対話について，wizard が提案したエンティティ名を user が受け入れた場合，wizard は前ターンの検索条件を変えてデータベースを再検索する必要はないため，検索クエリは更新されない．しかしそのエンティティ名はユーザの（暗黙的な）要求であるため，対話状態に含められるべきである．

対話状態アノテーションを付与するため，我々はクラウドワーカを募集し，検索クエリに含まれなかった非明示的な値をアノテーションした．各ワーカは，我々が用意したアノテーション用の UI を用いて，各ターンの対話状態をアノテーションした．アノテーションの質を担保するため，作業は練習と本番の二段階に分けて実施した．まず各ワーカは，アノテーション作業のマニュアルを読んだ後，練習として 10 対話分の作業を実施した．誤りのあった作業者には著者からの指摘・フィードバックが与えられ，再度練習を実施した．誤りの無くなった作業者のみが本番のアノテーション作業に参加した．

最終的に，6 名のクラウドワーカが分担して，全 4,246 対話に含まれる合計 30,593 ターン分の対話状態アノテーションを付与した．結果として，合計 58,745 件のスロット（検索クエリに記録されていたスロット数の約 37.8%）が対話状態の要素として追加された．以降では，この対話状態アノテーションが付与された JMultiWOZ を単に JMultiWOZ と呼ぶ．

4 ベンチマーク

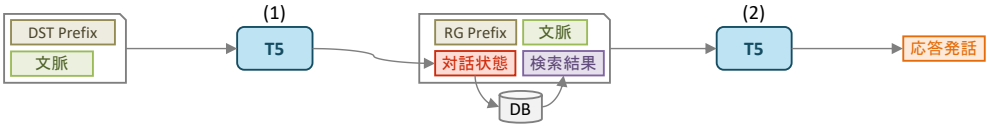
JMultiWOZ は，タスク指向型対話システムに必要となる 2 つのタスク，すなわち対話状態追跡 (DST) と応答生成 (RG) のベンチマークを提供する．DST は各ターンの対話状態を推定するタスクであり，RG は各ターンまでの対話履歴からシステムの次の応答を生成するタスクである．JMultiWOZ が，既存の英語対話データセットと同程度のベンチマークを提供できることを実証するため，MultiWOZ2.2[2] [21] における SOTA 手法と最新の LLM ベースの手法を用いて，上記 2 タスクを評価した．

4.1 ベースラインモデル

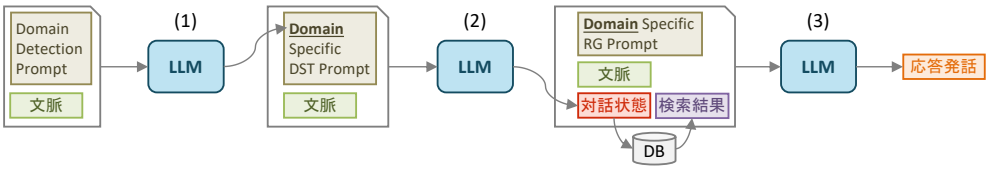
近年のタスク指向型対話モデルの構築手法は，中規模の事前学習済み言語モデルを対話データセットで fine-tuning する手法 [11] と，LLM を用い zero-shot もしくは few-shot 設定で応答生成する手法 [22] の 2 つに分類される．本研究では，JMultiWOZ に基づいて以下のように両手法を実装し評価した：

Fine-tuning MultiWOZ2.2 における SOTA モデルである，T5 [23] ベースのモデル TOATOD [18] を使用した．TOATOD の基盤モデルとしては多数のタ

2) MultiWOZ2.2 は，オリジナルの MultiWOZ における多数のアノテーションエラーを修正したバージョンである．



(a) T5 パイプライン [18]．(1) 対話履歴から対話状態が推定され，(2) その結果に基づいて最終的な応答が生成される．



(b) Zero-shot 設定での LLM パイプライン [19]．(1) 対話履歴から現在話題となっているドメイン（アクティブドメイン）を推定する．そして，アクティブドメインのみに着目して (2) 対話状態追跡と (3) 応答生成を実施する．

図 2　T5 および LLM を用いた，タスク指向型対話システムのパイプライン

スク指向型対話データセットで事前学習された T5 モデルが用いられるが，そのような資源は日本語では存在しない．そのため，ここでは，単に日本語データで事前学習された一般的な T5 モデル 2 つ (T5-base/large[3]) を基盤モデルとして使用する．

T5 の fine-tuning では，DST および RG の入出力のマッピングを，seq-to-seq にモデル化する．つまり，各ターンに対し，DST では対話履歴から対話状態へ，RG では対話履歴，対話状態，データベース検索結果から応答文へのマッピングをそれぞれ学習する．モデルの入出力，及び学習詳細は A.1 節を参照されたい．両タスクは単一の T5 によって同時に学習する．学習後の T5 による，DST と RG からなる end-to-end 応答生成のパイプラインを図 2a に示す．

LLM-based zero-/few-shot　Zero-/few-shot 設定における DST および RG には，LLM パイプライン [19] を用いる．図 2b に，zero-shot 応答生成のための 3 ステップからなるパイプライン処理を示す．Few-shot 設定においては，各ステップで用いられるプロンプトに，2 つの対話事例を含める．対話事例は，train セットから，対話文脈の埋め込み[4]が類似した対話を抽出することで得られる．

Zero-/few-shot のための LLM としては，OpenAI が API を提供する GPT-3.5 (`gpt-3.5-turbo`) と GPT-4 (`gpt-4`) を用いた．GPT-3.5 は，Hudeček ら [19] の評価実験内で検証された複数の LLM の中で，当時の最高性能を達成している．また GPT-4 は，日本語性能が最も高い LLM である[5]．プロンプトとしては，Hudeček ら [19] の研究で用いられたプロンプトを参考にしつつ，JMultiWOZ 向けに日本語で作成し直したものを用いた（A.2 節にプロンプト例を示す）．

3)　https://huggingface.co/retrieva-jp/t5-large-long
4)　日本語文埋め込みモデル (https://huggingface.co/cl-nagoya/sup-simcse-ja-large) を利用した．
5)　2023 年 10 月時点での日本語 LLM リーダボード (https://api.wandb.ai/links/wandb-japan/6ff86bp3) に基づく．

表 2　自動評価ベンチマーク結果．ダガー † は，Bang ら [18] が報告した，MultiWOZ2.2 における TOATOD (T5-base に相当) の評価値を引用していることを示す．ダブルダガー ‡ は，Hudeček ら [19] が報告した，MultiWOZ2.2 における GPT-3.5 の評価値を引用していることを示す．

Model	Few-shot	MultiWOZ 2.2			JMultiWOZ		
		JGA	Slot-F1	BLEU	JGA	Slot-F1	BLEU
T5-base†	✗	0.64	0.94	17.04	0.59	0.95	42.31
T5-large	✗	–	–	–	0.77	0.98	49.68
GPT-3.5‡	✗	0.13	0.40	4.17	0.16	0.70	5.31
GPT-4	✗	–	–	–	0.31	0.86	8.87
GPT-3.5‡	✓	0.27	0.51	6.77	0.25	0.82	12.91
GPT-4	✓	–	–	–	0.36	0.89	15.76

4.2　評価尺度

DST の評価尺度としては，joint goal accuracy (JGA) と Slot-F1 を用いた．JGA は，各ターンに推定された対話状態と，真の対話状態が完全一致したかどうかの 0/1 によって評価される．Slot-F1 は，各ターン推定された対話状態と真の対話状態の一致率を F1 によって評価される．RG の評価尺度としては，生成された応答文と真の応答文との BLEU を用いた．

4.3　結果

表 2 は，MultiWOZ2.2 における結果と JMultiWOZ における結果の比較を示している．JMultiWOZ では，DST と RG の両方において，fine-tuning 手法，すなわち，T5-base/large が最高性能であり，zero-/few-shot における LLM の性能には限界があった．この傾向は，MultiWOZ の結果と類似している．

表 3 人間評価実験結果．“N” は，各モデルと対話した評価者の数を示す．“Und.”，“App.”，“Sat.” は，それぞれ，システムの理解能力，システム応答の適切さ，対話の満足度のスコアを示す．† は，Iizuka ら [24] が報告した，MultiWOZ2.2 における各モデルの評価値を引用していることを示す．また**太字**は，MultiWOZ と比較して JMultiWOZ で大きく低下した Success スコアを示す．

Model	MultiWOZ 2.2						JMultiWOZ					
	N	Success	Turn	Und.	App.	Sat.	N	Success	Turn	Und.	App.	Sat.
T5-base†	36	66.67	12.56	3.83	3.81	3.72	38	65.79	10.74	3.92	3.71	3.55
T5-large	–	–	–	–	–	–	40	75.00	10.10	4.05	3.98	3.82
GPT-3.5†	42	57.14	11.55	3.79	3.98	4.05	41	**24.39**	11.05	2.90	2.24	1.95
GPT-4†	42	76.19	11.88	4.26	4.36	4.00	42	**57.14**	9.55	3.93	3.52	3.02

DST の標準的な尺度である JGA は，MultiWOZ と JMultiWOZ で大きな差は見られなかった．これは，JMultiWOZ の対話が，MultiWOZ のそれと同程度の複雑さを持ち，また，対話状態アノテーションの質が同程度であることを示している．したがって，JMultiWOZ は，既存データセットと同程度の DST ベンチマークを提供できることが示唆された．

RG の尺度である BLEU については，MultiWOZ に比べ JMultiWOZ が高い．この要因として，JMultiWOZ の対話では wizard の発話に一貫性がある，という点が考えられる．MultiWOZ とは異なり，JMultiWOZ の収集では，一つの対話に対し一人の wizard のみが従事した．さらに，wizard は対話収集前に十分な訓練を受けた．これら質の管理によってシステム発話の一貫性が達成されたと推察される．

5 人間評価実験

タスク指向型対話モデルのタスク達成能力を評価するためには，自動評価だけでなく，人間との実際の対話を用いた評価も重要である．本節では，4 節で用いた 4 つの対話モデルの end-to-end な対話能力を，クラウドワーカとの対話によって評価した．なお，GPT-3.5 および GPT-4 では，その性能の高さ（表 2 を参照）から few-shot 設定を用いた．

5.1 実験設定

本実験では，MultiWOZ において，TOATOD [18] と LLM パイプライン [19] をクラウドソーシングを用いた対話によって評価した Iizuka ら [24] の実験設定にならった．具体的には，各ワーカは，まずランダムな対話ゴールを提示され，そのゴールを達成できるように 4 つのシステムのうちいずれかと対話した．各対話は最大 20 ターンとし，ゴールが達成できたか（Success）をワーカが判断した．ワーカは対話終了後，システムの理解能力，システム応答の適切さ，対話の満足度のそれぞれを 5 段階で主観的に評価した．

5.2 結果

表 3 に，JMultiWOZ における結果，そして比較のため，Iizuka ら [24] が報告した MultiWOZ における結果をそれぞれ示す．Fine-tuning 手法である T5-base の Success は，MultiWOZ においてそれに対応する TOATOD（T5-base）のそれと大きな違いは見られなかった．したがって，JMultiWOZ を用いることで，日本語において，MultiWOZ と同程度の能力のタスク指向型対話モデルの学習・システム構築が可能であることが示唆された．

表 3 から，few-shot 設定における GPT-3.5 と GPT-4 の性能は，MultiWOZ の場合に比べ若干低下することがわかる．4 節におけるターン単位のスコアは低下していなかった（表 2 を参照）ことから，GPT-4 など高性能な LLM であっても，複数ターンからなる日本語対話における few-shot 対話生成能力は，英語と比較すると限界があると考えられる．

6 おわりに

本研究では，日本語における DST および RG のベンチマークを提供するため，JMultiWOZ に対して対話状態アノテーションを追加した．そして，MultiWOZ2.2 で検証されてきた fine-tuning ベースの対話モデルと，最新の LLM ベースのモデルを用いて，JMultiWOZ の DST および RG を学習し，本データセットが MultiWOZ と同難易度のベンチマークを提供できることを実証した．さらに，これら対話モデルの人間評価実験を実施し，最新の LLM であっても日本語におけるタスク指向型対話の能力には課題があることを示した．JMultiWOZ の活用によって，日本語におけるタスク指向型対話システムの研究開発が進展することを期待したい．

謝辞

本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けたものです．

参考文献

[1] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. In **Proc. NeurIPS**, pp. 20179–20191, 2020.

[2] Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. GALAXY: A Generative Pre-trained Model for Task-Oriented Dialog with Semi-supervised Learning and Explicit Policy Injection. In **Proc. AAAI**, No. 10, pp. 10749–10757, 2022.

[3] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. Recent advances and challenges in task-oriented dialog systems. **Science China Technological Sciences**, pp. 1–17, 2020.

[4] Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. End-to-end Task-oriented Dialogue: A Survey of Tasks, Methods, and Future Directions. In **Proc. EMNLP**, pp. 5925–5941, 2023.

[5] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-Value Retrieval Networks for Task-Oriented Dialogue. In **Proc. SIGDIAL**, pp. 37–49, 2017.

[6] Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a Conversational Agent Overnight with Dialogue Self-Play. **arXiv preprint arXiv:1801.04871**, 2018.

[7] Pawe l Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In **Proc. EMNLP**, pp. 5016–5026, 2018.

[8] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. In **Proc. AAAI**, Vol. 34, pp. 8689–8696, 2020.

[9] Johannes EM Mosig, Shikib Mehri, and Thomas Kober. STAR: A Schema-Guided Dialog Dataset for Transfer Learning. **arXiv preprint arXiv:2010.11853**, 2020.

[10] Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In **Proc. NAACL-HLT**, pp. 3002–3017, 2021.

[11] Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. Multi-task pre-training for plug-and-play task-oriented dialogue system. In **Proc. ACL**, pp. 4661–4676, 2022.

[12] Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. LAD: Language Models as Data for Zero-Shot Dialog. In **Proc. SIGDIAL**, pp. 595–604, 2022.

[13] Jeffrey Zhao, Yuan Cao, Raghav Gupta, Harrison Lee, Abhinav Rastogi, Mingqiu Wang, Hagen Soltau, Izhak Shafran, and Yonghui Wu. AnyTOD: A programmable task-oriented dialog system. In **Proc. EMNLP**, pp. 16189–16204, 2023.

[14] Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 281–295, 2020.

[15] Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In **Proc. EMNLP**, pp. 930–940, 2020.

[16] Yinpei Dai, Wanwei He, Bowen Li, Yuchuan Wu, Zheng Cao, Zhongqi An, Jian Sun, and Yongbin Li. CGoDial: A large-scale benchmark for Chinese goal-oriented dialog evaluation. In **Proc. EMNLP**, pp. 4097–4111, 2022.

[17] 大橋厚元, 平井龍, 飯塚慎也, 東中竜一郎. JMultiWOZ: 日本語タスク指向型対話データセットの構築. 言語処理学会 第 29 回年次大会, 2023.

[18] Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. Task-Optimized Adapters for an End-to-End Task-Oriented Dialogue System. In **Findings of ACL 2023**, pp. 7355–7369.

[19] Vojtěch Hudeček and Ondrej Dusek. Are Large Language Models All You Need for Task-Oriented Dialogue? In **Proc. SIGDIAL**, pp. 216–228, 2023.

[20] J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. **ACM Trans. Inf. Syst.**, Vol. 2, pp. 26–41, 1984.

[21] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In **Proc. NLP4ConvAI**, pp. 109–117, 2020.

[22] Gonçalo Raposo, Luisa Coheur, and Bruno Martins. Prompting, retrieval, training: An exploration of different approaches for task-oriented dialogue generation. In **Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 400–412, 2023.

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **JMLR**, pp. 1–67, 2020.

[24] Shinya Iizuka, Shota Mochizuka, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. Clarifying the Dialogue-Level Performance of GPT-3.5 and GPT-4 in Task-Oriented and Non-Task-Oriented Dialogue Systems. In **Proc. AI-HRI**, 2023.

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **Proc. ICLR**, 2019.

A Appendix

A.1 T5 モデルの fine-tuning 詳細

T5-base および T5-large の学習では，バッチサイズを 32 に設定し，5 エポック学習した．AdamW [25] オプティマイザを使用し，学習率は初期値 5e-5 からステップ数に応じて線形減衰させた．評価では最終ステップにおけるチェックポイントを使用し，DST 及び RG の推論では，いずれも greedy search を採用した．図 3 に，T5 の学習に使用された入出力系列の具体例を示す．

入力系列 対話から信念状態を推定: <顧客> 東京へ旅行に行くのですが、Wi-Fiが無料で利用できて、レストラン併設の宿泊施設はありますか？ <店員> WiFi無料でレストラン併設の宿泊施設は東京には１８件あります <顧客> 安めのところもありますか？ <店員> 値段が安めの条件を追加すると、残念ながら１件もヒットしません。 <顧客> そうですか。じゃあWi-Fiはなくてもいいです。
出力系列 general active_domain hotel, general city 東京, hotel pricerange 安め, hotel withrestaurant 有り

(a) DST における入出力データ

入力系列 対話から応答を生成: <顧客> 東京へ旅行に行くのですが、Wi-Fiが無料で利用できて、レストラン併設の宿泊施設はありますか？ <店員> WiFi無料でレストラン併設の宿泊施設は東京には１８件あります <顧客> 安めのところもありますか？ <店員> 値段が安めの条件を追加すると、残念ながら１件もヒットしません。 <顧客> そうですか。じゃあWi-Fiはなくてもいいです。 <信念状態> general active_domain hotel, general city 東京, hotel pricerange 安め, hotel withrestaurant 有り <検索結果> total 1, candidate [ビジネスホテル山百合], selected [city 東京, name ビジネスホテル山百合, genre ビジネスホテル, area 台東区, pricerange 安め, station [上野駅, 京成上野駅], wifi 無し, parking 無し, withrestaurant 有り, phone 0338317759, address 東京都...
出力系列 値段が安めでレストラン併設となれば、ビジネスホテル山百合、これ一択です。

(b) RG における入出力データ

図 3 T5 の DST タスクおよび RG タスクの学習に用いられた入出力データ具体例．入力系列の先頭には，DST と RG のいずれのタスクであるかを示す prefix が付与されている．また，入力系列の各要素には，話者を示す prefix や，信念状態，DB 検索結果を示す prefix が付与されている．

A.2 プロンプトの具体例

図 4 に，実験で LLM に入力された zero-shot 設定でのプロンプトを示す．

```
対話文脈から判断できる，レストランに関するスロットと値ペア（信念状態）を抽出してください．
信念状態は半角スペースとカンマを使用し，`スロット1 値1, スロット2 値2` という形式で抽出してください．
抽出するべきスロットは以下の通りです：
- "name" レストランの名前
- "genre" レストランのジャンル
- "area" レストランのエリア．"~区" や "~市" などの地区名．
- "pricerange" レストランの価格帯．"安め/普通/高め" のいずれか．
- "station" レストランの最寄り駅
- "wifi" レストランのWi-Fiの有無．"有り(無料)/有り(有料)/無し" のいずれか．
- "parking" レストランの駐車場の有無．"有り(無料)/有り(有料)/無し" のいずれか．
- "people" レストランの予約人数
- "day" レストランの予約日
- "time" レストランの予約時間
上記以外の情報は抽出しないでください．
また，文脈中で言及されなかったスロットの値も抽出しないでください．
```

(a) 飲食店ドメインにおける DST 用プロンプト

```
あなたは顧客の要望に沿ったレストランを探し出し，予約をするアシスタントです．
データベースを用い，エリア，ジャンル，価格帯等からレストランを検索・予約することができます．
レストランを見つけたら，その名前，住所，電話番号，その他必要な情報など，顧客から尋ねられた情報を
提供してください．
予約が成功したら，その予約番号（ref）を提供してください
```

(b) 飲食店ドメインにおける RG 用プロンプト

図 4 Zero-shot 設定での LLM の DST タスクおよび RG タスクに用いられた，飲食店ドメイン用のプロンプト例