後処理ネットワークを用いた強化学習によるタスク指向型対話システムの最適化

大橋厚元 [1]　東中竜一郎 [1]
[1] 名古屋大学大学院 情報学研究科
ohashi.atsumoto@g.mbox.nagoya-u.ac.jp,　higashinaka@i.nagoya-u.ac.jp

概要

典型的なタスク指向型対話システムは，複数のモジュールが連結したパイプライン型の構成となっている．近年では，このパイプライン型システムの対話タスク能力を向上させるため，システムを構成する複数のモジュールを同時に学習することでシステム全体を最適化する手法が多く提案されている．しかしこれら手法では，各モジュールが学習可能な手法で実装されていなければならない．本研究では，強化学習を用い，各モジュールの実装手法に依存せずにパイプライン型システム全体を同時に最適化できる後処理ネットワークを用いた手法を提案する．

1　はじめに

タスク指向型対話システムは複数のモジュールが連なって構成されるパイプライン型システムと End-to-End 型システムの 2 つに分類される [1, 2]．典型的なパイプライン型システムは，言語理解（Natural Language Understanding; NLU），状態更新（Dialogue State Tracking; DST），行動決定（Policy），言語生成（Natural Language Generation; NLG）という 4 つのモジュールで構成され [1]，各モジュールは人手で作成したルールによる手法（ルールベース手法）やニューラルネットワークを用いた手法（ニューラルベース手法）など，任意の手法で実装することができる．パイプライン型システムは，システム内部の各モジュールの入出力が明確であるので，開発者がシステムを解釈しやすいという利点がある．しかし，各モジュールが順番に処理されるため，先行するモジュールのエラーが後続するモジュールに伝播しやすく，結果として対話タスク能力が低下してしまうという欠点がある．

そこで，パイプライン型システムを構成する複数のモジュールを同時に学習することでシステム全体を End-to-End に最適化する手法が多く提案されている [3, 4, 5, 6]．しかしこれらの手法は，ニューラルベース手法による学習可能なモジュールでシステムが構成されていることを前提としている．したがって，ルールベース手法や Web API 等を用いて実装された学習不可能なモジュールを使用したい場面では，これら手法は適用することができない．

このような背景から，本研究ではシステムに含まれるモジュールがどのような手法で実装されていたとしても，システム全体を最適化できる後処理ネットワークを用いた手法を提案する．本手法では，システムを構成する各モジュールの出力を後処理するネットワーク（Post-processing Networks; PPN）を用意する．各 PPN は，後処理として各モジュールの出力を修正し，システム全体のパイプライン処理を円滑にする．各 PPN の後処理は，強化学習を用いてタスク達成などの対話タスク能力を向上できるように学習される．実際に学習されるのは各 PPN であるため，各モジュール自身が学習可能である必要はない．MultiWOZ データセット [7] に基づいて実装された複数のパイプライン型システムに PPN を適用し，本手法の有効性を確認した．

2　提案手法

本手法の目的は，PPN の後処理によって各モジュールの出力を修正することで，システム全体の対話性能を向上させることである．図 1 は，複数のモジュール（Module_A〜Module_N）からなるパイプライン型システムに PPN を適用した際の全体像を示している．

2.1　後処理アルゴリズム

図 1 を用いて，PPN_I がターン t における Module_I の出力 o_I^t を後処理するアルゴリズムを説明する．

まず，一般的なパイプライン型システムと同様

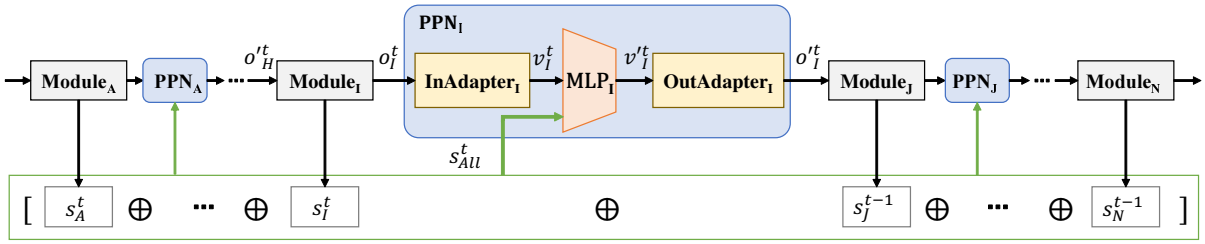

図 1 PPN を適用したパイプライン型システムのアーキテクチャ．各 PPN は対象とするモジュールの出力 o を，全モジュールの状態 s_{All} を考慮しながら後処理する．

に，Module_I は先行する Module_H の出力を受け取り，自身の処理結果 o_I^t を出力する．またこのとき，Module_I が自身の状態に関する補足的な情報（例えば，o_I^t の信頼度スコア）を出力できる場合はそれを状態ベクトル s_I として出力する．そしてこれを用いて，全モジュールの状態ベクトルを結合した $s_{All}^t = [s_A^t; ...; s_I^t; s_J^{t-1}; ...; s_N^{t-1}]$ を作成する．

次に，o_I^t を PPN_I に入力する．PPN_I 内では，$\text{InAdapter}_\text{I}$ が o_I^t のバイナリベクトル表現 v_I^t を作成する．v_I^t の次元数は Module_I の出力語彙数であり，各バイナリは「Module_I の各出力語彙が o_I^t に含まれているか否か」を 1 か 0 で表す．$\text{InAdapter}_\text{I}$ は，Module_I に定義されている出力語彙セットに基づいて人手で作成する．

ここまでで作成された v_I^t と s_{All}^t を結合し，多層パーセプトロン MLP_I に入力する．MLP_I は v_I^t と同じ次元数のバイナリベクトル $v_I''^t$ を出力する．この時点での v^t と v''^t の変化分が PPN_I による後処理となる．最後に $\text{OutAdapter}_\text{I}$ によって，$v_I''^t$ の中で 1 となっている要素が Module_I の出力語彙に復元され，最終的な PPN の出力 $o_I''^t$ が作成される．$\text{OutAdapter}_\text{I}$ も $\text{InAdapter}_\text{I}$ と同様に，Module_I に定義されている出力語彙セットに基づいて人手で作成する．

2.2 強化学習による最適化

本研究では対話シミュレーションを用い，システムに含まれる各 PPN の MLP を方策ネットワークとして強化学習を行う．強化学習で用いる報酬としては，対話タスクの達成に関するものを人手によるルールで設計する．強化学習アルゴリズムとしては，方策ベースの手法であり学習の安定性も高いという理由から，Proximal Policy Optimization（PPO）[8] を用いた．以下は，PPO に基づいた PPN の学習ループ 1 回分の手順である．この学習ループは事前に決められた回数繰り返す．

1. ユーザシミュレータを用いて対話をシミュレーションする．対話における各ターン t では，各 PPN が後処理を行い，そこで得られる s_{All}^t，v^t，v''^t，報酬 r^t からなるタプルを行動履歴として蓄積する．対話シミュレーションは行動履歴が事前に決められた量（ホライゾンと呼ぶ）に達するまで繰り返す．
2. **PPN 選択戦略**（後述）に基づいて，今回の学習ループで更新する PPN を選択する．
3. ステップ 1 で得られた行動履歴を用い，事前に決められたエポック数だけ PPO アルゴリズムに基づいて，ステップ 2 で選択した PPN の MLP を更新する．

PPN 選択戦略とは，各学習ループで更新する PPN を選択するためのルールである．システムに 2 つ以上の PPN が含まれる場合，どの PPN をどの順番で更新するべきかは自明ではない．そこで本研究では **ALL**（各学習ループで，必ずすべての PPN を選択する），**RANDOM**（各学習ループで，1 つ以上の PPN をランダムに選択する），**ROTATION**（学習ループごとに，PPN を 1 つずつ順繰りに選択する）の 3 種類を用意し，最も有効な戦略を実験的に調査する．

3 実験

3.1 データセットとプラットフォーム

本研究では，MultiWOZ データセット [7] に基づいて実装されたモジュールを用いて複数の異なる対話システムを構成し，各システムに PPN を適用することで，本手法の有効性を検証する．MultiWOZ は旅行サポートサービスを行う店員と顧客とのタスク指向型対話コーパスであり，10,438 対話が含まれる．

実験では，タスク指向型対話システム用プラットフォームである ConvLab2 [9] を用いる．ConvLab2 では，MultiWOZ に基づいて実装された各モジュールやユーザシミュレータ [10] を用いた対話システム評価ツール等が提供されている．

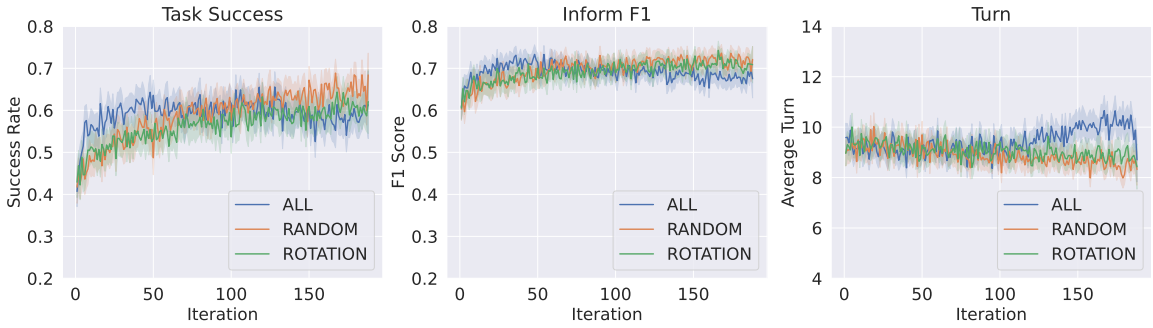

図 2　強化学習における PPN 選択戦略ごとの各評価尺度の推移

3.2　評価尺度

各対話におけるタスク評価には，ConvLab2 で利用可能であり，先行研究 [11, 12, 13] でも一般的に使用されている，**Turn**（対話に要したターン数），**Inform F1**（ユーザが要求した情報に過不足なく答えたか），**Match Rate**（ユーザの条件に合ったエンティティ[1)]を提示したか），**Task Success**（20 ターン以内に Inform Recall と Match Rate が同時に 1 になったか）の 4 つの尺度を用いる．なお本実験では，ユーザ発話とそれに対するシステムの応答 1 回を 1 ターンとして定義する．システムの評価時には公正な比較を行うため，異なる 5 種類のランダムシードで 1,000 対話ずつシミュレーションし，その平均スコアを用いる．

3.3　システム構成

本研究では，Takanobu ら [14] の実験に倣い，パイプライン型システムの一般的なモジュール構成である，NLU，DST，Policy，NLG の 4 モジュールからなるシステムを用意する．各モジュールとしては，ConvLab2 で実装されているモデル[2)]の中から，古典的なルールベース手法（学習不可能なモデル）と最新のニューラルベース手法（学習可能なモデル）の両方を含むように選択した．なお，学習可能なモデルであってもそのモデル自体は学習せず PPN のみを学習する．以下で，モジュール別に，今回利用するモデルを説明する．

NLU　BERT[15] NLU モデルを使用する．本モデルは BERT による埋め込み表現を用い，ユーザ発話文中の各単語が表す意図をクラス分類する [16]．

1）　エンティティとは，○○レストランや××ホテルといった，MultiWOZ のデータベースに含まれる施設のことである．
2）　各モデルとしては，ConvLab2 で提供されている 2021 年 10 月 20 日時点でのベストモデルを選択した．

表 1　強化学習後の PPN 選択戦略ごとの各スコア

PPN 選択戦略	Task Success	Inform F1	Match Rate	Turn
ALL	64.2	**71.9**	76.6	9.20
RANDOM	**66.1**	71.5	**78.7**	**8.61**
ROTATION	60.4	70.5	73.2	9.10

DST　Rule DST と TRADE [17] の 2 モデルを使用する．Rule DST は NLU が推定したスロットを用いて対話状態を更新するモデルであり，更新ルールは人手によるルールで作成されている．一方で TRADE はニューラルベース手法のモデルであり，システムとユーザの発話履歴を入力として，対話状態を直接生成する．

Policy　Rule Policy，MLE Policy，PPO [8] Policy，LaRL [18] の 4 モデルを使用する．Rule Policy は人手で作成されたルールに従って行動を決定するモデルである．MLE Policy は MultiWOZ データセットにおける店員の行動を模倣するように教師あり学習されたニューラルベースモデルである．PPO Policy は MLE Policy をベースに PPO アルゴリズム [8] によってタスク達成に最適化するように fine-tune されたモデルである．LaRL は対話状態を元にシステムの発話文を直接生成する RNN ベースのモデルである．

NLG　Template NLG と SCLSTM [19] の 2 モデルを使用する．Template NLG は，人手で事前に用意されている発話テンプレート文の集合から，Policy が出力した行動に合うものを選択する手法である．SCLSTM は RNN ベースのモデルであり，Policy が出力した行動に合ったシステム発話文を生成する．

3.4　結果

本研究では，まずどの PPN 選択戦略が最も有効であるかを調査し，次に PPN が任意のシステムに適用可能であるかを調査した．

表 2　各システムにおけるモデルの組み合わせと PPN 適用前後での評価尺度の比較

System	Model Combination				Use PPN	Task Success	Inform F1	Match Rate	Turn
	NLU	DST	Policy	NLG					
SYS-RUL	BERT	Rule	Rule	Template	—	84.1	87.4	90.2	5.92
					✓	84.0	86.3	**92.4**	6.33
SYS-MLE	BERT	Rule	MLE	Template	—	43.3	62.4	27.8	9.03
					✓	**66.1**	**71.5**	**78.7**	**8.61**
SYS-PPO	BERT	Rule	PPO	Template	—	54.9	65.5	55.2	8.41
					✓	**68.8**	**72.1**	**77.8**	**8.37**
SYS-SCL	BERT	Rule	Rule	SCLSTM	—	38.3	57.5	56.7	13.53
					✓	**44.2**	**71.7**	**71.8**	**11.04**
SYS-TRA	TRADE		Rule	Template	—	19.0	45.6	36.4	12.08
					✓	18.8	**49.2**	31.6	12.14
SYS-LAR	BERT	Rule	LaRL		—	21.6	44.9	27.6	13.24
					✓	**23.9**	**50.9**	**34.1**	**12.77**

3.4.1　PPN 選択戦略の比較

図 2 と表 1 は各 PPN 選択戦略を用いてシステムを学習した場合の学習推移と最終スコアをそれぞれ示している．なおこの実験では BERT NLU，Rule DST，MLE Policy，Template NLG の組み合わせで構成したシステムに PPN を適用した．このシステムを用いた理由は，PPN 適用前の Task Success が 50% 前後（図 2 左図参照）であるため PPN の適用による影響がわかりやすく，また MLE Policy はしばしば強化学習の初期重みとして用いられる [11, 12] ため，本研究における初期モデルとして合理的だと考えられるからである．結果から，いずれの PPN 選択戦略であっても PPN が有効であることが確認できる．なお，ALL はほかの 2 種類に比べ早く最高スコアに到達するが，それ以降では学習が不安定になることが分かる．一方 RANDOM と ROTATION は学習が安定しており，特に RANDOM は最終的に全 3 種類中最高スコアを達成している．

3.4.2　モデルの組み合わせによる影響

表 2 は複数のモデル（3.3 節参照）を用いて実装した 6 種類のシステムの構成と PPN 適用前後での評価尺度を示している．PPN 選択戦略としては，3.4.1 節の結果から RANDOM を用いた．表 2 から，ほとんどのシステムにおいて，PPN を適用することによってその対話タスク能力が改善することがわかる．一方で特に SYS-RUL など，Rule Policy が含まれるシステムについては改善が見られない場合も確認できる．Rule Policy は人手ルールによって入念に作りこまれており元から精度が高いモデルである．したがって，特に Policy に改善の余地がある場合に PPN による改善が見込めると考えられる．

4　関連研究

本研究は，モジュール構造をとるシステムを End-to-End で最適化するニューラルベース手法と関連している．Lei ら [20] は DST と NLG に相当する 2 つの decoder を sequence-to-sequence モデル [21] に導入した手法を提案している．さらに Policy や NLU に相当する decoder によって Lei らのモデルを拡張した手法も提案されている [4, 22]．しかし，これら手法は End-to-End モデルを教師あり学習により最適化しており，対話全体の教師データ必要とする点が本研究とは異なる．

本研究は，強化学習によるパイプライン型システムの性能向上を目的としている．Liu ら [5] は模倣学習を組み合わせることで，Policy を実ユーザに最適化させる手法を提案している．また，システム内の各モジュールを同時に fine-tune するすることで，システム全体の対話能力やエラーに対するロバスト性を向上させる手法 [23, 24, 6] が多く提案されている．これら手法は，ニューラルベース手法で実装された各モジュールを強化学習によって直接学習するため，学習不可能なモジュールを含むシステムには適用できない．一方で本研究で提案する手法は，任意のパイプライン型システムに適用可能である．

5　おわりに

本研究では後処理ネットワークと強化学習を用い，各モジュールの実装手法に依存せずにパイプライン型対話システム全体を同時に最適化できる手法を提案した．MultiWOZ と ConvLab2 における実験から，多様なモデルで構成された複数のシステムに対して本手法が有効であることを確認した．

謝辞

本研究は科研費「モジュール連動に基づく対話システム基盤技術の構築」（課題番号 19H05692）の支援を受けた.

参考文献

[1] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. **ACM SIGKDD Explorations Newsletter**, 19(2):25–35, 2017.

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

[3] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A Network-based End-to-End Trainable Task-oriented Dialogue System. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pages 438–449, April 2017.

[4] Yichi Zhang, Zhijian Ou, and Zhou Yu. Task-Oriented Dialog Systems That Consider Multiple Appropriate Responses under the Same Context. In **Proceedings of the AAAI Conference on Artificial Intelligence**, pages 9604–9611, Apr. 2020.

[5] Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pages 2060–2069, June 2018.

[6] Zichuan Lin, Jing Huang, Bowen Zhou, Xiaodong He, and Tengyu Ma. Joint System-Wise Optimization for Pipeline Goal-Oriented Dialog System. **arXiv preprint arXiv:2106.04835**, 2021.

[7] Paweł 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 **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pages 5016–5026, October-November 2018.

[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. **arXiv preprint arXiv:1707.06347**, 2017.

[9] Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pages 142–149, July 2020.

[10] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In **Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers**, pages 149–152, April 2007.

[11] Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. Guided Dialog Policy Learning: Reward Estimation for Multi-Domain Task-Oriented Dialog. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pages 100–110, November 2019.

[12] Ziming Li, Sungjin Lee, Baolin Peng, Jinchao Li, Julia Kiseleva, Maarten de Rijke, Shahin Shayandeh, and Jianfeng Gao. Guided Dialogue Policy Learning without Adversarial Learning in the Loop. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pages 2308–2317, November 2020.

[13] Zhengxu Hou, Bang Liu, Ruihui Zhao, Zijing Ou, Yafei Liu, Xi Chen, and Yefeng Zheng. Imperfect also Deserves Reward: Multi-Level and Sequential Reward Modeling for Better Dialog Management. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pages 2993–3001, June 2021.

[14] Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. In **Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pages 297–310, July 2020.

[15] 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, Volume 1 (Long and Short Papers)**, pages 4171–4186, June 2019.

[16] Qian Chen, Zhu Zhuo, and Wen Wang. Bert for joint intent classification and slot filling. **arXiv preprint arXiv:1902.10909**, 2019.

[17] Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pages 808–819, July 2019.

[18] Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pages 1208–1218, June 2019.

[19] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pages 1711–1721, September 2015.

[20] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 1437–1447, July 2018.

[21] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In **Proceedings of Advances in neural information processing systems**, pages 3104–3112, 2014.

[22] Weixin Liang, Youzhi Tian, Chengcai Chen, and Zhou Yu. MOSS: End-to-End Dialog System Framework with Modular Supervision. **Proceedings of the AAAI Conference on Artificial Intelligence**, 34(05):8327–8335, Apr. 2020.

[23] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-End Task-Completion Neural Dialogue Systems. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pages 733–743, November 2017.

[24] Hwaran Lee, Seokhwan Jo, Hyungjun Kim, Sangkeun Jung, and Tae-Yoon Kim. SUMBT+LaRL: Effective Multi-Domain End-to-End Neural Task-Oriented Dialog System. **IEEE Access**, 9:116133–116146, 2021.