タスク指向対話における強化学習を用いた対話方策学習への敵対的学習の役割の解明

下山翔 [1,2]　森村哲郎 [3]　阿部拳之 [3]
[1]AI Shift, Inc.　[2] 東北大学大学院 理学研究科 数学専攻　[3]CyberAgent, Inc. AI Lab.
shimoyama_sho@ai-shift.jp　{morimura_tetsuro, abe_kenshi}@cyberagent.co.jp

概要

強化学習を用いた対話方策の学習において，報酬関数の設計は重要である．効果的な報酬関数の人手による設計が困難であるため，データから報酬関数を推定する手法の 1 つとして敵対的逆強化学習が注目されている．この方法により推定された報酬関数を用いて学習された対話方策は良好な性能を示すことが報告されている．この理由を明らかにするため，本研究では，既存手法の 1 つである Guided Dialog Policy Learning に対する分析を行う．報酬関数の目的関数に対する考察および実験結果から，敵対的学習はループ対話の発生を抑制する効果を持つことを示す．

1　はじめに

対話方策はタスク指向対話システムの重要な構成要素の一つである [1, 2]．近年，強化学習を用いた対話方策の学習に関する研究が注目されている [3, 4, 5, 6, 7]．強化学習による対話方策の学習は，対話をマルコフ決定過程とみなし，対話方策と環境（例えば，ユーザシミュレータ [8, 9]）との相互作用を通じて，得られる対話に対する報酬を最大化することで行われる．

報酬関数は対話方策の学習において重要である一方，適切な報酬関数の人手による設計は困難である．単純な報酬関数として，対話ターンを短くするために各対話ターンで一定の小さな負の報酬，対話終了時に目的を達成した場合は大きな正の報酬を与える報酬関数があげられる．しかしながら，各ターンで選択された行動の良さは対話が終了するまで分からない（つまり，報酬が疎である）ため対話方策の学習が適切に進まない場合が存在する．

上述の問題に対処するため，データから報酬関数を推定する方法の 1 つである敵対的逆強化学習 [10, 11] を用いた方法が提案されている．研究 [12] は，エキスパート対話を用い，対話単位の報酬関数と対話方策を敵対的学習を用い交互に学習する方法を提案している．研究 [13] は敵対的逆強化学習を用いて報酬関数を推定し，報酬関数を状態行動単位とすることで報酬が疎な問題に対処する手法 Guided Dialog Policy Learning (GDPL) を提案している．これらの報酬関数を用いて学習された対話方策は人手で設計した報酬関数を用いて学習された対話方策より良好な性能を示すことが報告されている．しかし，対話方策の学習において，敵対的学習を用いた報酬関数の推定が有用であることの理由は十分に解明されていない．

本研究では，GDPL を対象とし，強化学習を用いた対話方策の学習において，敵対的学習を用いた報酬関数推定が対話方策の学習に与える影響の分析を行う．報酬関数の目的関数に対し考察を行い，実験によりその妥当性の評価を行う．

2　関連研究

強化学習における人手での報酬関数の設計が困難であるため，データから報酬関数を推定する方法である逆強化学習 [14] がよく用いられる．特に，近年，逆強化学習に敵対的学習を用いた手法である敵対的逆強化学習 [10, 11] が注目されている．研究 [12] は対話方策の学習と報酬関数の学習を敵対的学習を用いて交互に行う手法を提案している．報酬関数を識別器とし，エキスパート対話には高い値を，対話方策が生成した対話には低い値を割り当てるように学習する．対話方策は報酬関数から得られる報酬が高くなるよう学習される．研究 [13] は対話方策と状態行動単位の報酬関数を交互に学習する手法である Guided Dialog Policy Learning (GDPL) を提案している．GDPL は各ターン毎に状態行動に対して報酬を与えることで報酬が疎となる問題に対処し

ており，対話内における話題の転換を柔軟に捉えることができることが報告されている．研究 [15] は報酬関数の推定と対話方策の学習を交互ではなく逐次的に行う手法を提案している．敵対的学習を用い，状態行動単位の報酬関数である識別器とノイズを入力とし状態行動対を出力する生成器を学習する．次に，上記で学習した報酬関数を固定し，対話方策を学習する．報酬関数の学習と対話方策の学習を分けることで，方策オフ型 / オン型どちらにも適用でき，モード崩壊 [16] に対処している．先行研究と本研究の違いは敵対的学習が対話方策の学習に及ぼす影響を詳細に分析している点である．

3 前提

本論文では，対話を離散時間有限状態マルコフ連鎖 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \pi_\theta, \nu, r, \gamma, T)$ として扱う．ここで，$\mathcal{S}$ は dialog state [17, 18] 全体からなる集合（状態集合），$\mathcal{A}$ は dialog act [19] 全体からなる集合（行動集合），$P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ は状態遷移確率（$\Delta(\mathcal{S})$ は $\mathcal{S}$ 上の確率単体），$\pi_\theta : \mathcal{S} \to \Delta(\mathcal{A})$ は対話方策（$\Delta(\mathcal{A})$ は $\mathcal{A}$ 上の確率単体，θ はパラメータ），$\nu \in \Delta(\mathcal{S})$ は初期分布，r は報酬関数，$\gamma \in [0, 1]$ は減衰定数，$T \in \mathbb{N}$ は最大ターン数を表す．$s_t \in \mathcal{S}, a_t \in \mathcal{A}$ でそれぞれターン $t \in \{1, \ldots, T\}$ における状態，行動，$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{|\tau|}, a_{|\tau|}\}$ ($|\tau| \leq T$) で対話を表す．

3.1 Guided Dialog Policy Learning

Guided Dialog Policy Learning (GDPL) [13] は，敵対的学習を用い，報酬関数と対話方策を交互に学習する手法である．対話方策が生成した対話に対して低い報酬，学習データ（例えば人間同士の対話）内の対話に対して高い報酬を割り当てるように敵対的学習を用いて報酬関数を学習する．対話方策は報酬関数から得られる報酬が高くなるように学習することで，人間に近い対話の実現を目指す．GDPL では，状態 s 行動 a の報酬を $f_\omega(s, a) = \log p_\omega(s, a)$（ここで，$p_\omega(s, a)$ は確率モデル，ω はそのパラメータ）と表し，以下の目的関数 (1), (2) を用いて対話方策 $\pi_\theta(a|s)$ と報酬関数 f_ω を交互に更新する：

$$\max_\theta \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=t_0}^{|\tau|} \gamma^{t-t_0} \hat{r}(s_t, a_t) \right], \tag{1}$$

$$\min_\omega \left\{ \mathrm{KL}[p_\mathcal{D} | p_\omega] - \mathrm{KL}[p_{\pi_\theta} | p_\omega] \right\}, \tag{2}$$

ここで，$\hat{r} = f_\omega - \log \pi_\theta$, $p_\mathcal{D}(s, a)$ は経験分布，p_{π_θ} は π_θ に従った際に状態行動 (s, a) が発生する同時確率分布，$\mathrm{KL}[q_1|q_2] = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_1(s, a) \log \frac{q_1(s,a)}{q_2(s,a)}$ である．GDPL では，p_ω を推定する代わりに確率であるという制約なしで f_ω を直接推定し，対話方策の学習を安定させるために potential-based reward shaping [20] を用いて (s_t, a_t, s_{t+1}) に依存する形に f_ω を変換している．

4 GDPL に対する考察

本章では，GDPL において，敵対的学習が対話方策の学習に与える影響の考察を行う．4.1 節でループ対話の定義，4.2 節で敵対的学習の役割の考察を行い，敵対的学習はループ対話を抑制する効果を有することを示す．

4.1 ループ対話の定義

表 1 ループ対話の例．usr はユーザシミュレータ，sys は対話方策を表す．

話者	発話
usr	Need a restaurant called meze bar restaurant.
sys	There 's a place called meze bar restaurant.
	I will book it for you and get a reference number ?
usr	I am looking for details on the meze bar restaurant restaurant.
sys	Would you like to try meze bar restaurant ?
usr	I am looking for details on the meze bar restaurant restaurant.
sys	Would you like to try meze bar restaurant ?
	...
usr	I am looking for details on the meze bar restaurant restaurant.
sys	Would you like to try meze bar restaurant ?

Definition 1 (ループ対話)**.** 対話 τ において，$(s_t, a_t) = (s_{t+n}, a_{t+n})$ を満たす n とターン t が存在するとき，τ をループ対話と呼び，ループ内の状態 $s^i := s_{t+i}$ for $\forall i \in \{0, \ldots, n-1\}$ をループ状態，行動 $a^i := a_{t+i}$ for $\forall i \in \{0, \ldots, n-1\}$ をループ行動と呼ぶ．

ループ対話の例を表 1 に示す．この対話では上から 2 つ目のユーザ発話（$t = 1$）以降，ユーザシミュレータと対話方策がともに同じ内容の発話を繰り返している．簡単に述べると，ループ対話とは，特定の状態行動 (s, a) が複数回出現する対話である．

4.2 敵対的学習の役割の考察

マルコフ連鎖 $\mathcal{M}$ において，ある $\{(s^i, a^i)\}_{i=0}^{n-1}$ が存在し，次の仮定を満たすとする：任意の $i \in \{0, \ldots, n-1\}$ に対し $\hat{r}(s^i, a^i)$ が $\hat{r}$ の最大値と十分近い．このとき，GDPL において，敵対的学習はループ対話の発生を抑制すると推測される．

報酬関数を固定し，式 (1) に基づき対話方策を更新する場合を考える．式 (1) は累積報酬であるから，対話方策は生成する対話の累積報酬が高くなるよう学習される．ある 2 つの対話に対し，それらの対話に含まれる状態行動に対する報酬の差が十分小さい場合，累積報酬が高くなる対話はターン数が多い対話である．ここで，そのような対話の 1 つとして，$\{(s^i, a^i)\}_{i=0}^{n-1}$ をループ状態行動に持ち，あるターン t 以降これらの状態を順に遷移し続けるループ対話が考えられる（つまり，$\tau = \{s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s^0, a^0, \ldots, s^{n-1}, a^{n-1}, s^0, a^0, \ldots\}$）．したがって，式 (1) を最大化して得られる対話方策の 1 つとして，上述のループ対話の発生を促す対話方策が考えられる．(s^i, a^i) が 1 対話内に複数回出現することから，このような対話方策に従った際の (s^i, a^i) の発生確率 $p_{\pi_\theta}(s^i, a^i)$ は十分大きいと考えられる．

上述したループ対話の発生を促す対話方策 π_θ を固定し，式 (2) に基づき報酬関数を推定する場合を考える．式 (2) は以下のように書ける：

$$\max_{\omega} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} f_\omega(s, a)\{p_{\mathcal{D}}(s, a) - p_{\pi_\theta}(s, a)\}.$$

上式において，$p_{\mathcal{D}}(s, a) < p_{\pi_\theta}(s, a)$ なら，$f_\omega(s, a) \to -\infty$ である．ここで，(s, a) をループ状態行動 (s^i, a^i) とすると，π_θ はループ対話の発生を促すという仮定から $p_{\pi_\theta}(s^i, a^i)$ は十分大きいと考えられる．よって，$p_{\mathcal{D}}(s^i, a^i) < p_{\pi_\theta}(s^i, a^i)$ を満たし，$f_\omega(s^i, a^i) \to -\infty$ となると考えられる．したがって，このように更新された報酬関数を用いて対話方策を更新する場合，(s^i, a^i) を含む対話は累積報酬が十分小さくなることから，対話方策は (s^i, a^i) の発生確率 $p_{\pi_\theta}(s^i, a^i)$ が低くなるように学習されると考えられる．つまり，GDPL において敵対的学習はループ対話を抑制すると考えられる．

5 考察に対する実験的確認

本章では，4 章における考察に関する実験的確認を行う．確認する内容は以下である：

- GDPL から敵対的学習を取り除いて学習した対話方策はループ対話の発生を促す．
- GDPL において敵対的学習はループ対話を抑制する．
- 敵対的学習を用いずにループ対話を抑制するよう報酬を設計した手法と GDPL の性能差．

5.1 データセットとシミュレータ

本研究では，データセットとして Multiwoz [21] を用いる．Multiwoz はマルチドメインタスク指向対話コーパスで，7 ドメイン，10,483 対話からなる．

強化学習の環境として，対話ゴールを受け取り確率的なルールに基づいて対話を行うアジェンダベースユーザシミュレータ [8] を用いた．対話ゴールは，各ドメインに対する，シミュレータが希望するエンティティが満たすべき制約（例えば，適度な価格帯のレストラン）と対話方策から聞く必要がある要求情報（例えば，レストランの住所）から成る．

5.2 比較手法

本論文では，以下 2 つの方法でそれぞれ学習した対話方策を GDPL と比較する：

- GDPL w/o AL: 式 (2) から第 2 項を取り除き学習した f_ω を固定し，$\hat{r}$ を式 (3) で定め，式 (1) を目的関数として学習した対話方策．
- GDPL restrict loop: GDPL w/o AL と同様の f_ω を固定し，$\hat{r}$ を式 (4) で定め，式 (1) を目的関数として学習した対話方策．

$$\hat{r}(s_t, a_t, s_{t+1}) = \frac{f_\omega(s_t, a_t, s_{t+1})}{10000} - \log \pi_\theta(a_t|s_t) \tag{3}$$

$$\hat{r}(s_t, a_t, s_{t+1}) = \begin{cases} r_{\text{loop}} - \log \pi_\theta(a_t|s_t) & ((s_t, a_t) \text{ is loop state action}) \\ \frac{f_\omega(s_t, a_t, s_{t+1})}{10000} - \log \pi_\theta(a_t|s_t) & (\text{otherwise}) \end{cases} \tag{4}$$

$$r_{\text{loop}} = \min_{(s_t, a_t, s_{t+1}) \in \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{valid}}} \left\{ -\frac{|f_\omega(s_t, a_t, s_{t+1})|}{2} \right\}$$

ここで，$\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{valid}}$ はそれぞれ学習，検証データである．GDPL w/o AL において，学習を安定させるために，報酬関数を 10000 で割っているが割らない場合においても同様の結果を確認した．

GDPL の実装には著者らのコード[1)]を用いた．$T = 40$ とし，モデルのパラメータおよび dialog state, dialog act は文献 [13] と同様とした．

5.3 評価指標と評価・学習方法

評価指標として success を用いた．success は recall と match rate がともに 1 なら 1，そうでないなら 0 として定める．recall はシミュレータのゴールで指定

1） https://github.com/truthless11/GDPL

された要求情報と実際に対話方策が伝えた情報を用いて計算する．match rate は予約されたエンティティがシミュレータのゴールで指定された制約と完全に一致していたドメインの割合である．

全データの内，8483 対話を学習，1000 対話を検証に用いた．シードをランダムに 5 つ変更し，各手法をそれぞれ 50 エポック学習した．ここで，1 エポックは 1024 対話（つまり，環境との相互作用 1024 セッション）とした．各手法間でシードは同じものを用いた．

各エポック終了時に，ユーザシミュレータとの相互作用により評価用対話を 1000 生成し，それらの対話に対する評価指標の平均を報告する．評価対話の生成に用いるシードは全てのエポック，手法で共通とした．

5.4 結果と分析

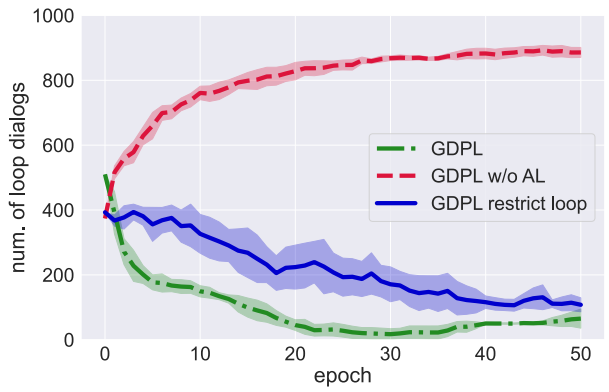

図 1 各報酬の与え方に対するループ対話数の推移．GDPL w/o AL は GDPL における報酬関数の目的関数から敵対的学習項を取り除いたもの，GDPL restrict loop は GDPL w/o AL においてループ状態行動に対し負の報酬を与えたものを表す．

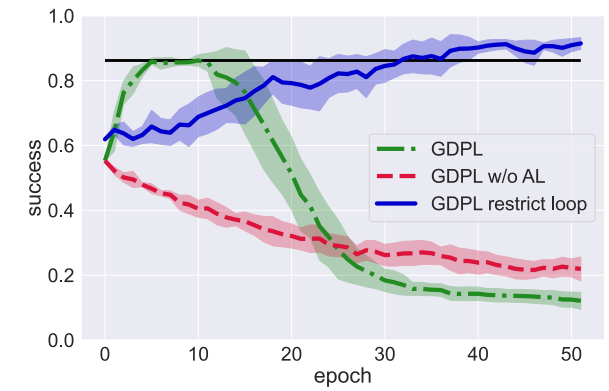

図 2 各報酬の与え方に対する success rate の推移．黒の横線は GDPL に対する平均値の最大値を表す．

評価用対話 1000 対話に対するループ対話数の平均と ± 標準偏差を図 1 に示す．GDPL w/o AL において，ループ対話数が増加している．一方，GDPL においてループ対話数は減少している．よって，GDPL において，敵対的学習はループ対話の発生を抑制すると考えられる．

評価用対話 1000 対話に対する各評価指標の平均と ± 標準偏差を図 2 に示す．GDPL および GDPL restrict loop は GDPL w/o AL より良好な性能を示している．この結果から，タスク指向対話において，ループを抑制することが対話方策の性能向上に寄与すると推測される．加えて，GDPL restric loop は GDPL より良好な性能を示している．10 エポック目以降 GDPL の性能が悪化しているが，これは no-offer で終了する対話が増加していることが要因であった．no-offer は dialog-act の値の 1 つであり，ユーザシミュレータは no-offer を含んだ dialog-act を受け取った時点で対話を終了する．no-offer により対話が終了すると，それ以降で聞くはずであった情報が聞けず，また，それ以降で予約するはずであったエンティティが予約できないため評価指標が低下する．no-offer を含む対話を学習データから取り除き GDPL を学習したところ，学習途中で性能が悪化しなかったことから，この現象は敵対的学習のモード崩壊 [16] であると考えられる．なお，モード崩壊とは，生成モデルの学習に失敗し出力のバリエーションが限られる現象である．GDPL restrict loop は学習に敵対的学習を用いていないため，GDPL におけるモード崩壊の問題を回避できていると考えられる．

6 結論と将来の展望

本研究では，Guided Dialog Policy Learning (GDPL) において，敵対的学習を用いた報酬関数の推定が対話方策の学習に与える影響の分析を行った．報酬関数の目的関数を状態行動について書き下すことで，敵対的学習はループ対話の発生を抑制する効果を有すると推測された．また，アジェンダベースユーザシミュレータを用いて，上述の考察の実験的確認を行った．実験において，敵対的学習を用いずにループ対話の発生を抑制するよう報酬を与えた方法は GDPL より良好な性能を示した．今後の課題として次を挙げる：アジェンダベース以外のシミュレータ（例えば，Variational Hierarchical User Simulator [9]）を用いた実験的分析および敵対的学習を用いた他手法への本分析の適用可能性の調査．

謝辞

本論文の作成にあたりご助力頂きました, 株式会社 AI Shift の友松祐太氏，杉山雅和氏，東佑樹氏，二宮大空氏，戸田隆道氏，株式会社サイバーエージェントの邊土名朝飛氏にこの場を借りて厚く御礼申し上げます.

参考文献

[1] Wai-Chung Kwan, Hongru Wang, Huimin Wang, and Kam-Fai Wong. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning, 2022.

[2] 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*, pp. 142–149, 2020.

[3] Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. Policy networks with two-stage training for dialogue systems. In *In 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 101–110, 2016.

[4] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

[5] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. Deep dyna-q: Integrating planning for task-completion dialogue policy learning, 2018.

[6] Huimin Wang, Baolin Peng, and Kam-Fai Wong. Learning efficient dialogue policy from demonstrations through shaping. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6355–6365, 2020.

[7] Huimin Wang and Kam-Fai Wong. A collaborative multi-agent reinforcement learning framework for dialog action decomposition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7882–7889, 2021.

[8] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 149–152, 2007.

[9] Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. User modeling for task oriented dialogues. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 900–906, 2018.

[10] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.

[11] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[12] Bing Liu and Ian Lane. Adversarial learning of task-oriented neural dialog models. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 350–359, 2018.

[13] 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*, pp. 100–110, 2019.

[14] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. p. 1433–1438, 2008.

[15] 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*, pp. 2308–2317, 2020.

[16] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, 2017.

[17] J. Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue Discourse*, pp. 4–33, 2016.

[18] Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 567–582, 2020.

[19] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, pp. 339–374, 2000.

[20] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, p. 278–287, 1999.

[21] 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, 2018.