情報検索に基づく応答復元とのマルチタスク学習に基づく長期間対話のための応答生成

Response Generation for Long-term Conversation via Multi-Task Learning with IR-based Response Restoration

髙﨑環 [1*]　吉永直樹 [2]　豊田正史 [2]
Meguru Takasaki[1]　Naoki Yoshinaga[2]　Masashi Toyoda[2]

[1] 東京大学大学院情報理工学系研究科
[1] Graduate School of Information Science and Technology, the University of Tokyo
[2] 東京大学生産技術研究所
[2] Institute of Industrial Science, the University of Tokyo

Abstract: When a dialogue system has a long-term conversation with a person, it is desirable to generate responses taking past dialogue sessions into account. However, the conversation logs used for training dialogue systems do not necessarily contain many responses considering the past dialogue context. Therefore, it is difficult to generate responses that fully respect the past dialogue context if the dialogue system is only trained by concatenating the past dialogue context with the current context. In this paper, we propose a multi-task learning method for response generation to force the dialogue system to consider the past context adequately. The auxiliary self-supervised task is to generate the system-side utterance included in the most similar past dialogue context to the current context. In the experiment, we trained our proposed models on the Mulit-session Twitter Dialogue Dataset and verified the effect of our data augmentation methods.

1　はじめに

スマートフォンやスマートスピーカーの普及に伴い，知的対話アシスタント（例: Apple Siri, Amazon Alexa）とユーザが継続的に会話を行う機会が増えている．このような対話システムとの会話においてユーザのエンゲージメントを高めるためには，特定のタスクを実行するための会話だけでなく，雑談にも対応することが重要である [2]．多様な話題を扱う雑談に対応するためには，SNS 上の大規模な対話ログを用いて応答生成モデルを学習することが有効である [18]．多くの対話システムでは直前の会話のみを参照し応答生成を学習するが，過去のやりとりを考慮した人間の会話とは異なり，応答の自由度が高すぎる [20] ことが課題として挙げられる．そこで，対話システムと特定のユーザが継続的に対話することを想定し，過去の対話セッションの情報を参照し応答生成する手法が提案されている [23, 24, 1, 21]．

しかし，豊富な対話履歴を参照可能であっても，過去の対話セッションの情報を十分に尊重した応答は生成されづらいと予想される．その理由の一つとして，応答生成モデルの学習に用いる会話ログには過去の対話文脈を踏まえた応答が必ずしも多くはなく，現在文脈に追加入力してモデルを学習するだけでは，過去の対話を考慮した応答生成が難しいことが考えられる．

そこで本研究では，過去文脈をより参照した応答生成を行うためのマルチタスク学習手法を提案する．本手法では，検索された過去文脈を用いた応答生成タスクだけでなく，現在文脈と検索された過去文脈から最も類似した過去文脈の応答部分を抽出するタスクを解く事例を追加し，生成モデルの学習を行う．追加タスクを解く事例は，応答生成データセットにおける過去文脈の検索結果から構築可能で，追加のアノテーションが不要である．また，応答生成を解く事例と過去文脈を抽出する事例の入出力の形式は一致しているため，アーキテクチャの増強を必要としない．

実験では，Twitter 長期間対話データセット [21] を用いて学習・推論を行い，提案するデータ拡張手法によって訓練されたモデルの応答性能を検証した．自動評価の結果，学習タスクの追加によって応答性能が改善されることがわかった．

*連絡先：東京大学生産技術研究所
〒 153-8505 東京都目黒区駒場 4 丁目 6 - 1
E-mail: takasa-m@tkl.iis.u-tokyo.ac.jp

2 関連研究

2.1 長期間の文脈を考慮した応答生成

近年，複数の対話セッションの履歴を考慮した雑談応答生成手法として，アーキテクチャを拡張して階層的に処理する手法 [22, 26] や，過去の対話文脈を圧縮する手法 [23, 24, 1]，そして過去の対話文脈を部分的に抽出する手法 [23, 21] が提案されている．

Wu ら [22] は過去の対話文脈をターンごとに埋め込み，動的にメモリを更新しながら応答生成に用いる手法を提案した．また Zhang ら [26] は，過去の対話セッションの埋め込み表現と使用語彙を参照しながら応答を生成する手法を提案している．

一方，過去の対話文脈を圧縮し応答生成モデルに入力する手法が提案されている．Xu ら [23] は，過去の対話文脈から生成した要約を参照しつつ，応答を生成する手法を提案した．また Xu ら [24] は，ユーザとチャットボット双方のプロフィール文を保持・参照・更新を行い，応答を生成する手法を提案した．これらの研究を基に Bae ら [1] は，対話要約を動的に更新しつつ，要約を用いて応答を生成する手法を提案した．

また，追加の教師データが不要な方法として，過去の対話文脈を抽出し応答生成に用いる手法が提案されている．Xu ら [23] は現在文脈と類似した過去の対話セッションを参照し応答生成する方法を提案した．Takasaki ら [21] はより細かい単位で過去の対話文脈を検索・抽出することで，Xu らの応答生成手法を改善した．

本研究では，過去の対話文脈を抽出する手法 [21] を基に，検索結果に基づくデータ拡張によるマルチタスク学習手法を提案する．

2.2 複数発話文脈を強く参照した応答生成

現在進行中の対話のみを参照する応答生成では，文脈への摂動に対して生成応答の変化が少なく [19]，発話単位での関係性を十分に考慮した応答生成が難しいと考えられている．この課題に対し，会話の流れを考慮した学習 [6, 11]，摂動を用いた学習 [27, 28]，そして事前学習済みモデルの微調整・推論に特化した工夫 [7, 14] を行う手法が提案されている．

発話同士の時系列的関係を考慮した応答生成のために，Hao ら [6] は文脈間の類似性を考慮するモジュールを追加し学習を行う手法を提案した．一方 Li ら [11] は，近接発話との潜在表現の差分を考慮し，会話の流れを大域的に捉えるための事前学習手法を提案した．

さらに，対話文脈への敵対的摂動を用いた目的関数の設計も提案されている．Zhao ら [27] は，モデルが文脈により注意を向けるように，トークン・発話レベルでの文脈への摂動を復元するタスクを加える学習手法を提案した．一方 Zhou ら [28] は敵対的摂動に対し生成応答がより変化するように報酬を設計し，多様で一貫した応答生成を学習する手法を提案した．

また，モデルの増強や目的関数における工夫とは異なり，事前学習済みモデルを微調整・生成する際の手法も提案されている．He ら [7] は事前学習に使用したデータを使用し破滅的忘却を緩和する微調整手法を提案し，文脈への摂動に敏感な応答生成を可能にすると主張した．また，Malkin ら [14] は事前学習済みモデルを微調整後，入力する文脈の長さを変えた際の予測結果を組み合わせることで，文脈を考慮した多様な応答を生成できる手法を提案した．

一方本研究では長期間にわたる過去の対話履歴から抽出された文脈を入力としており，入力系列上で距離が近い文脈同士が時系列的にも隣接しているとは限らず，文脈の時間的隣接性を考慮する手法は有効でないと考えられる．また，理想的には過去文脈の一部を現在セッションと同時に考慮しながら応答生成を行うべきであるため，セッション内で完結する従来の摂動復元タスクだけでは効果的な学習が難しいと予想される．そこで本研究では，現在文脈と過去文脈を同時に参照するために，現在文脈に最も類似した過去文脈中の応答を抽出するタスクを追加するマルチタスク学習手法を提案する．

2.3 タグ付きデータによるマルチドメイン学習

学習データ内に異なるドメインのテキストが混在する際，事例ごとにドメインを示すタグを付与して学習を行うことで，言語処理タスクにおいて性能を向上させる手法が研究されている．Johnson ら [8] は，複数のターゲット言語が存在する機械翻訳タスクにおいて，入力文の先頭に翻訳先言語を示すタグを付与して学習を行うことで，単一のモデルで多言語データを翻訳することを可能とした．また Caswell ら [4] は，逆翻訳によって拡張された対訳データと正規の対訳データを区別するタグを入力文頭に付与し翻訳モデルを学習させることで，翻訳性能を向上させた．一方 Britz ら [3] は，出力文頭にドメインを識別するタグを付与し，翻訳とドメイン予測を同時に学習することによって，単一ドメインで学習した翻訳モデルを上回る翻訳性能を可能とした．

本研究では，過去文脈の応答を抽出するタスクを応答生成と同時に学習する際に，各事例にタスクを識別するためのタグを付与して学習する．タスクごとにモジュールを追加するマルチタスク学習手法 [27] とは異なり，事例追加によって様々な入出力に対応可能であり，追加のアーキテクチャ不要で学習可能である．

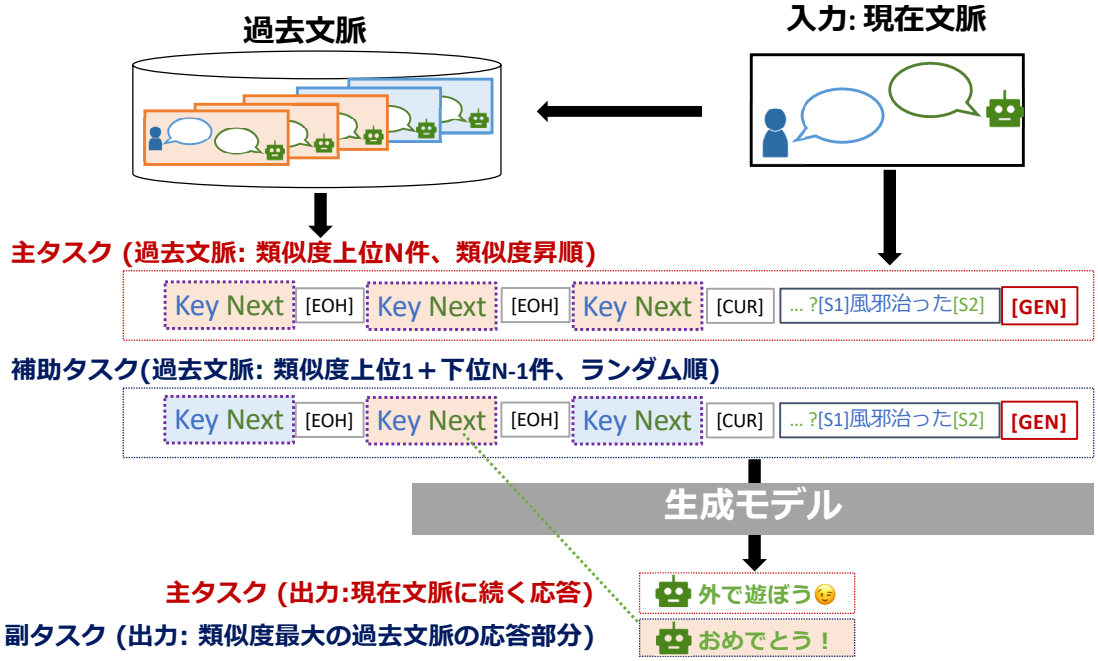

図 1: 提案するマルチタスク学習手法の外観（value を KeyNext とした場合）．主タスク（応答生成）では現在の対話文脈と，類似過去文脈を入力し，応答を生成する．補助タスク（応答抽出）では，現在の対話文脈と，最も類似した過去文脈 1 つ，類似しない過去文脈を複数入力し，類似過去文脈の応答部分を生成する．

3 提案手法

本研究では，過去の対話文脈を十分に考慮した応答を生成するために，応答生成より過去文脈の参照が必要なタスクを解く事例を追加するマルチタスク学習手法を提案する．本研究で追加するタスクでは，現在文脈と複数の過去文脈を入力とし，現在文脈と最も類似した過去文脈の応答部分を抽出するように学習を行う．提案手法の概観を図 1 に示す．

3.1 応答生成タスクの設計

応答生成を行う事例では，Takasaki ら [21] の手法をもとに，過去の対話文脈から現在文脈に類似した文脈を検索・抽出し，現在文脈と生成モデルに入力する．まず，現在/過去の対話文脈から同じ発話数で構成された query/key を作成し，それぞれの文脈を Sentence-BERT [17] を用いベクトル化する．そして，query に類似した key を選択したのち，key を対応する周辺文脈 (value) に変換する．本実験では，Takasaki ら [21] の手法で最も性能が高かった value として，応答生成モデルの最大入力長が 256 トークンの時は key に後続する発話 (Next), 最大入力長が 1024 トークンの場合は key と後続発話 (KeyNext) を採用した．value に変換された過去の対話文脈は，現在文脈とともに応答生成器に入力し，続く応答を予測するように学習・推論を行う．

3.2 応答抽出タスクの設計とデータ拡張

過去文脈を参照する必要とするタスクとして，入力された過去文脈のうち最も現在文脈に類似した文脈の応答部分を生成することを提案する．この応答抽出タスクを解く事例は，応答生成事例と似た形式で入出力を作成可能であり，追加のモジュール・アノテーションを必要とせず，応答生成タスクを解くためのデータセットから作成可能である．さらに，応答生成における参照応答に類似した形式・内容を出力するように，3.1 で説明した value のうち，key の後続発話にあたる部分 (Next) を生成するようにする．

本実験では応答生成事例から同じ現在文脈を持つ応答抽出事例を作成することで，応答抽出事例を同数分獲得する．応答抽出タスクにおける正解文脈と不正解文脈の差異を大きくするために，現在文脈と類似度最大のものは 1 つだけ選択し，残りは類似度が低いものを複数選択する．そして，正解文脈の位置が固定されないように，選択された過去文脈をランダムに並べ替えたのち，現在文脈とともに応答生成モデルに入力される．この時，入力最大長を超過した過去文脈が切り捨てられないように，正解文脈は左端以外に配置した．

訓練手法	追加割合	応答生成 Perplexity (↓)	BLEU-2/3 (↑)	ROUGE-L (↑)	BERTScore (↑)	DISTINCT-1/2 (↑)	文脈生成 Perplexity (↓)
最大入力長: 256 トークン							
直近履歴入力 [23]		45.53	1.99/0.82	10.24	60.70	5.03/19.09	-
応答生成 [21]		**44.21**	2.14/0.88	10.65	60.95	5.31/20.46	-
+応答抽出	0%	44.39	2.31/1.09	10.25	60.20	5.04/19.39	9.02
	25%	44.39	2.57/1.13	10.98	61.43	**6.47**/**24.98**	1.77
	50%	44.65	**2.64**/**1.14**	**11.47**	**61.64**	5.70/22.95	**1.72**
	100%	45.35	2.38/1.07	10.92	61.21	6.14/23.95	**1.72**
最大入力長: 1024 トークン							
直近履歴入力 [23]		42.94	2.02/0.84	10.12	61.32	5.47/22.67	-
応答生成 [21]		**41.78**	2.10/0.86	10.42	61.59	6.20/25.39	-
+応答抽出	0%	42.52	**2.74**/**1.42**	10.45	60.84	5.38/21.33	10.58
	25%	42.36	2.49/1.16	10.70	61.55	**6.48**/**25.76**	1.32
	50%	43.02	2.64/1.08	**11.57**	**62.04**	5.71/25.01	1.25
	100%	43.61	2.71/1.30	11.17	61.49	6.14/24.52	**1.21**

表 1: 自動評価の結果.

3.3 タグ付きデータによる生成モデルの学習

生成モデルは GPT-2 [16] の重みを初期値として，応答生成タスクと応答抽出タスクを解く事例を混ぜ，パラメータは完全に共有した状態で微調整を行う．この時，事例が解くタスクを示すために，応答生成タスクの場合には `[GEN]` トークン，類似過去文脈生成タスクの場合には `[EXT]` トークンを入力末尾に付与する．いずれの事例の場合も，選択された過去の対話文脈をその末尾で現在文脈と結合し，モデルの最大入力長を超過した分だけ入力系列の左端から切り捨て，生成モデルに入力する．この時，発話の開始には話者を示す `[S1]`, `[S2]` トークン，発話末尾には `</s>` トークン，過去文脈の終端には `[EOH]` トークン，現在文脈と過去文脈の境界には `[CUR]` トークンを付与する．応答生成事例では参照応答を，応答抽出タスクでは最も類似した過去文脈の Next 部分を生成するように，クロスエントロピー損失を用いて学習を行う．

4 実験

本節では，過去文脈抽出器を統合した GPT-2 ベースの対話モデルを，Twitter 長期間対話データセット [21] を用いて訓練し，提案手法の応答性能を評価する．

4.1 モデル

モデルの実装には PyTorch 1.10.2[1] および Transformers 4.20.0[2] を用いた．過去文脈抽出器における Sentence-BERT [17] は，Taksasaki ら [21] の手法同様，Twitter 長期間対話データセット [21] を用い，日本語版 BERT [5][3] に Pooling 層を加え 5 エポック分訓練を行ったのち，開発データでの分類精度が最良のモデルを採用した．また応答生成モデルは，日本語版 GPT-2 [16][4] を patience を 1 とする early stopping を用いて最大 5 エポック分微調整し，開発データでの損失が最小となるモデルについてそれぞれ性能を評価する．この時，提案手法の訓練における開発データは，訓練データと同様タグ付きデータ拡張を行ったものを使用していることに注意されたい．

ベースラインとして，直前の過去セッションを時系列順に入力する手法 [23]，そして過去文脈を数発話で検索・抽出する手法 [21] で最良のモデル (query/key を 1 発話とし，最大入力長が 256 トークンでは Next，1024 トークンでは KeyNext を value とするモデル) を使用した．提案手法では，過去文脈を数発話で検索・抽出する手法 [21] を基に，応答生成だけでなく過去文脈の応答抽出を行う事例を追加し学習した．1 つの応答生成事例につき，同じ現在文脈を持つ応答抽出事例が 1 つ作成後，応答生成事例における参照応答の最大長を上回る事例を除外することで，データ拡張を行った．この時タスクの割合による性能変化を確認するために，追加データ数を応答抽出事例全体の 25，50，100%としてモデルを学習させた．また，ベースラインとの差分として，末尾のタスク識別トークンの有無とそれに伴う語彙の追加，そして訓練中の評価に用いる検証データにおける応答抽出事例の有無があげられる．参考として，文脈生成データを訓練データには追加しない一方，それ以外は提案手法同様の設定で学習したモデル（表 1 中の 0%）についても性能評価を行なった．

[1] `https://pytorch.org/`
[2] `https://huggingface.co/`
[3] `https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking`
[4] `https://huggingface.co/rinna/japanese-gpt2-small`

モデル	妥当性	文脈一貫性	人間らしさ
応答生成 [21]	48.10	45.22	49.44
+応答抽出 (25%追加)	**52.03**	**46.81**	**51.68**
(p 値)	0.004	0.047	0.046

表 2: 人手評価の結果.

4.2 自動評価

各手法の応答生成・類似過去文脈生成における性能を自動評価指標を用いて比較する．応答生成の評価指標としては，Perplexity, BLEU-2/3 [15], ROUGE-L [12], BERTScore [25]，DISTINCT-1/2 [10] を採用した．生成応答の単語分割には MeCab[5] (UniDic 2.1.2 辞書) を用い，BERTScore の重みには日本語版 RoBERTa [13][6] を使用した．また上記に加え，本手法で追加した類似過去文脈生成における Perplexity の計測も行なった．

生成応答を自動評価した結果を表 1 に示す．応答生成事例のみで学習するベースラインに比べ，過去文脈生成データを追加した場合（25%以上）では一貫して BLEU-2/3，ROUGE-L, BERTScore が改善しており，より参照応答に類似した応答が生成できていると言える．これは，類似した文脈に続く Next を生成するように学習したことで，過去文脈を無視するバイアスを軽減できたためだと考えられる．

次に，参照応答を生成する際の PPL は，文脈生成タスクを追加することで一貫して悪化する一方で，類似過去文脈を生成する際の PPL はデータを増やすにつれて改善している．これは，学習データ中のタスクの割合によって学習時に重視するタスクの重みが異なるためであると考えられる．

また，DISTINCT-1/2 の観点では，最大入力長が 256 トークンの場合の提案手法（value:Next）では一貫して改善されている．一方，最大入力長が 1024 トークンの場合の提案手法（value:KeyNext）では 25%追加の場合のみスコアが改善された．このことから，追加データの割合によって，学習されたモデルが生成する応答の多様性は変化しうると予想される．

4.3 人手評価

提案手法とベースラインの応答性能を人手評価により比較する．1 人のアノテータが手法ごと 116 個の生成応答に対し，以下 3 つの評価基準について 0-100 の整数値で点数をつけ，手法ごとの平均値を算出した．

妥当性: 直前の入力発話に対して妥当な応答か

[5] https://taku910.github.io/mecab/

[6] https://huggingface.co/nlp-waseda/roberta-base-japanese

訓練手法	追加割合	現在文脈 平均	現在文脈 標準偏差	過去文脈 平均	過去文脈 標準偏差
最大入力長: 256 トークン					
直近履歴入力 [23]		0.794	0.188	0.576	0.249
応答生成 [21]		0.785	0.192	0.636	0.235
+応答抽出	0%	0.770	0.203	0.619	0.232
	25%	0.804	0.181	0.656	0.234
	50%	0.796	0.188	0.669	0.236
	100%	0.799	0.187	0.730	0.195
最大入力長: 1024 トークン					
直近履歴入力 [23]		0.782	0.191	0.672	0.228
応答生成 [21]		0.775	0.191	0.673	0.222
+応答抽出	0%	0.757	0.207	0.678	0.222
	25%	0.784	0.194	0.694	0.222
	50%	0.781	0.189	0.728	0.216
	100%	0.785	0.185	0.732	0.217

表 3: 参照応答から文脈への Attention の強さの比較.

文脈一貫性:（過去の対話セッションを含む）対話文脈と矛盾しない応答か

人間らしさ: 人間らしい応答か

過去セッション数が平均 15.49 であることを考慮し，15-17 個の過去セッションを含む 10 エピソードを評価データセットからサンプリングし，最終 3 セッションの生成応答を評価した．ベースラインは過去文脈を数発話で検索・抽出する手法 [21] を，提案手法は応答抽出事例の追加割合が 25%，のものを採用した．また，全ての手法の最大入力長は 1024 トークンとした．

人手評価した平均値を図 2 に示す．応答抽出タスクを追加することで，全ての評価指標において平均値が向上していることがわかる．また，ブートストラップ法 [9]（標本データ数は 116 個，反復回数は 1000 回）での p 値は全ての基準で 0.05 を下回り，平均値の有意差が確認された．

4.4 過去文脈への Attention の分析

応答生成時の文脈への参照度を分析するために，訓練データにおいて参照応答から現在・過去文脈へ向けられる Attention を分析する．まず，文脈内の各トークンに対し，層数 × ヘッド数 × 参照応答のトークン数分の Attention の最大値を計算する．その後，各事例の文脈に含まれるトークンごとの Attention から最大値を算出することで，事例ごとの文脈参照度を算出する．

各手法における事例ごとの文脈参照度の平均，標準偏差を表 3 に示す．現在文脈に比べ過去文脈への参照は一貫して弱いものの，応答抽出の事例数を増やすにつれ，過去文脈への参照度が大きくなることがわかる．このことから，応答抽出事例の追加によって過去文脈をより参照した応答生成が学習できると考えられる．

5 まとめ

本研究では，過去の対話文脈を十分に考慮した応答を生成するために，対話システムの学習の際に，検索された過去の対話文脈の応答部分を抽出するタスクを加える手法を提案した．Twitter 長期間対話データセットを用いた実験の結果，提案するマルチタスク学習によって応答性能が改善されることを示した．今後の取り組むべき課題としては，文脈に対する摂動復元などの別タスクを追加した，より洗練された学習手法の考案が挙げられる．

謝辞

本研究は JST CREST JPMJCR19A4 および JSPS 科研費 JP21H03445, JP21H034 の支援の助成を受けたものです．

参考文献

[1] S. Bae, D. Kwak, S. Kang, M. Y. Lee, S. Kim, Y. Jeong, H. Kim, S.-W. Lee, W. Park, and N. Sung. Keep me updated! Memory management in long-term conversations. In *Findings of EMNLP*, 2022.

[2] T. W. Bickmore and R. W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Interact.*, 2005.

[3] D. Britz, Q. Le, and R. Pryzant. Effective domain mixing for neural machine translation. In *WMT*, 2017.

[4] I. Caswell, C. Chelba, and D. Grangier. Tagged back-translation. In *WMT*, 2019.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[6] C. Hao, L. Pang, Y. Lan, F. Sun, J. Guo, and X. Cheng. Ranking enhanced dialogue generation. In *CIKM*, 2020.

[7] T. He, J. Liu, K. Cho, M. Ott, B. Liu, J. Glass, and F. Peng. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *EACL*, 2021.

[8] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 2017.

[9] P. Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, 2004.

[10] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL: HLT*, 2016.

[11] Z. Li, J. Zhang, Z. Fei, Y. Feng, and J. Zhou. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *ACL and IJCNLP*, 2021.

[12] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL*, 2003.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[14] N. Malkin, Z. Wang, and N. Jojic. Coherence boosting: When your pretrained language model is not paying enough attention. In *ACL*, 2022.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019.

[17] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, 2019.

[18] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *EMNLP*, 2011.

[19] C. Sankar, S. Subramanian, C. Pal, S. Chandar, and Y. Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. In *ACL*, 2019.

[20] S. Sato, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa. Modeling situations in neural chat bots. In *ACL SRW*, 2017.

[21] M. Takasaki, N. Yoshinaga, and M. Toyoda. Effective Dialogue-Context Retriever for Long-Term Open-Domain Conversation. In *IWSDS*, 2023.

[22] Q. Wu and Z. Yu. Stateful memory-augmented transformers for dialogue modeling, 2022.

[23] J. Xu, A. Szlam, and J. Weston. Beyond goldfish memory: Long-term open-domain conversation. In *ACL*, 2022.

[24] X. Xu, Z. Gou, W. Wu, Z.-Y. Niu, H. Wu, H. Wang, and S. Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of ACL*, 2022.

[25] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi. BERTscore: Evaluating text generation with bert. In *ICLR*, 2020.

[26] T. Zhang, Y. Liu, B. Li, Z. Zeng, P. Wang, Y. You, C. Miao, and L. Cui. History-aware hierarchical transformer for multi-session open-domain dialogue system. In *Findings of EMNLP*, 2022.

[27] Y. Zhao, C. Xu, and W. Wu. Learning a simple and effective model for multi-turn response generation with auxiliary tasks. In *EMNLP*, 2020.

[28] W. Zhou, Q. Li, and C. Li. Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. In *ACL and IJCNLP*, 2021.

情報検索に基づく応答復元とのマルチタスク学習に基づく長期間対話のための応答生成: 正誤表

1　謝辞

1.1　誤

本研究は JST CREST JPMJCR19A4 および JSPS 科研費 JP21H03445, JP21H034 の支援の助成を受けたものです.

1.2　正

本研究は JST CREST JPMJCR19A4 および JSPS 科研費 JP21H03445, J21H03494 の支援の助成を受けたものです.