Vol.2021-NL-249 No.11
2021/7/28

間接的な応答と直接的な応答の対からなる対話コーパスの構築

高山 隼矢[1,a)]　梶原 智之[2,b)]　荒瀬 由紀[1,c)]

概要：人間は対話においてしばしば相手の質問や発話に対して間接的な応答をする．例えば，予約サービスにおいてユーザがオペレータに対して「あまり予算がないのですが」と応答した場合，オペレータはその応答には間接的に「もっと安い店を提示してください」という意図が含まれていると解釈することができる．大規模な対話コーパスを学習したニューラル対話モデルは流暢な応答を生成する能力を持つが，間接的な応答に焦点を当てたコーパスは存在せず，モデルが人間と同様に間接的な応答を扱うことができるかどうかは明らかではない．本研究では既存の対話コーパスである MultiWoZ を拡張し，間接的な応答と直接的な応答の対からなる 7 万件規模の対話コーパスを構築した．ユーザーからの入力発話を事前により直接的な発話に言い換えることで対話応答生成の性能が向上することを確認した．

1. はじめに

対話において人間はしばしば自身の要求や意図について直接的に言及せず，間接的に表現する [1]．人間は対話相手から間接的な応答を受け取ったとき，これまでの対話履歴などの文脈に基づいて言外の意図を推測できる．図 1 に，レストランの予約に関する対話における間接的な応答と直接的な応答の例を示す．この例ではオペレータの「A レストランを予約しますか？」という質問に対してユーザは「予算が少ないのですが」と応答している（図中の「間接的な応答」）．この応答は字義通りの意味だけを考慮するとオペレータの質問への直接的な回答にはなっていない．しかし，オペレータは対話履歴を考慮してユーザが A レストランよりも安いレストランを探していると推論し，新たに A レストランよりも安い B レストランを提案している．人間と自然なコミュニケーションを行う対話システムの実現のためには，ユーザの間接的な応答に暗示された意図（直接的な応答）を推定する技術の実現が重要である．

大規模な対話コーパス [2–4] と深層学習技術により，近年では対話応答生成 [5, 6] や対話状態追跡 [7, 8] など様々なタスクにおいて高い性能を誇る手法が提案されている．また，最近では間接的な応答に関するコーパス [9, 10] も構築されている．しかし，Pragst ら [9] のコーパスはルールベースで半自動的に構築されたコーパスであり，多様性や自然さに欠ける．また，Louis ら [10] のコーパスは Yes/No 型かつ一問一答型の質問応答対のみを扱うコーパスであり，それ以外の間接的な応答には対応できない．間接的な応答の意図を解釈する対話システムの実現のためには，より複雑かつ自然な間接的な応答を含む対話コーパスの構築が必要である．

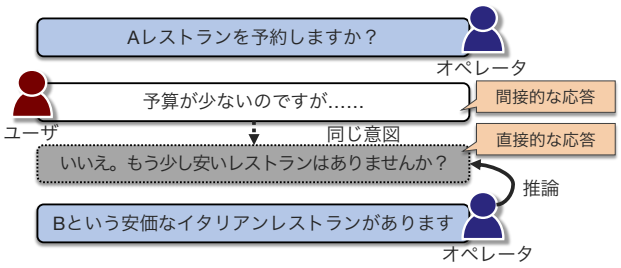

図 1 対話における間接的な応答と直接的な応答の例．これらの応答は字義通りに解釈すると異なる意味を持つが，この対話履歴上においては同じ意図の応答として解釈できる．

本研究ではより複雑な間接的な応答を理解する対話システムの開発のために，71,498 の間接的な応答と，その意図を明示的に表現した直接的な応答の対からなる対話コーパス DIRECT (Direct and Indirect REsponses in Conversational Text)[*1] を構築する．既存の対話システム研究の成

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[2] 愛媛大学大学院理工学研究科
Graduate School of Science and Engineering, Ehime University
[a)] takayama.junya@ist.osaka-u.ac.jp
[b)] kajiwara@cs.ehime-u.ac.jp
[c)] arase@ist.osaka-u.ac.jp

[*1] https://github.com/junya-takayama/DIRECT/

果の再利用を容易にするため，本コーパスは既存のマルチドメイン・マルチターンのタスク指向対話コーパス MultiWoZ [4] を拡張して作成する．具体的には，MultiWoZ の各ユーザ応答に対してクラウドソーシングを用いて「ユーザ応答をより間接的に言い換えた応答」と「ユーザ応答をより直接的に言い換えた応答」の対を収集する．そのため，DIRECT コーパスには元の応答・間接的な応答・直接的な応答の 3 つ組が収録される．

モデルが間接的な応答の意図を正確に解釈できるかどうかを検証するために，本研究では DIRECT コーパスを用いて，間接的な応答を直接的な応答へと言い換えるタスクを設計する．ベースラインとして最先端の事前学習済み言語モデルである BART [11] に基づく言い換えモデルを構築し，性能調査を行った．また，言い換えモデルを用いてユーザの入力発話を事前により直接的な発話に言い換えることで，対話応答生成の性能が向上することを確認した．

2. 関連研究

Pragst ら [9] は同じ意図を持つ間接的な応答と直接的な応答の対が含まれた対話コーパスをルールベースで自動構築する手法を提案している．また，再帰的ニューラルネットを用いた発話選択器を用いることで，対話中の間接的な応答をその応答と同じ意図を持つ直接的な応答に置き換えられることを示している．しかし，ルールベースであるために自動生成可能な対話データのパターンは限られており，多様性に欠けるコーパスとなっている．Louis ら [10] は，Yes/No 型の質問と，それに対する間接的な応答の対 34,268 件からなるコーパスを構築している．各質問応答対に対しては，応答が質問に対して肯定と否定のどちらの意図を示すものであるかがアノテーションされている．コーパス構築にあたってはまず著者らが事前に用意した 10 パターンの対話シナリオのいずれかに基づく質問をクラウドソーシングを用いて収集しており，質問数は合計で 3,431 件である．また，それぞれの質問に対して最大で 10 件の応答をクラウドソーシングを用いて収集している．そのため，人間らしくかつ多様な質問応答対からなるコーパスとなっている．しかし，一問一答型かつ Yes/No 型の質問応答対に限定されており，Yes/No で言い換えられないような間接的な応答には対応していない．本研究ではこれらの既存研究とは異なり，人手で作成された対話履歴に対して人手で間接的な応答と直接的な応答の対を作成する．そのため DIRECT コーパスは人間らしく多様な応答が含まれ，かつ複雑な間接的応答を扱う最初の対話コーパスである．

DIRECT コーパスにおける間接的応答と直接的応答の対はそれぞれ言い換え可能な関係にある．対話における発話の言い換え関係を扱った事例としては本研究の他に Hou ら [12] や Gao ら [13] が挙げられる．これらの研究では対話データ中の各発話に対し，その発話と同じ意図を持つ発話を言い換え生成技術を用いて複数パターン生成することでデータ拡張を行い，応答生成の性能向上を達成している．しかし，これらの研究は間接的な応答に着目したものではない．

表 1 間接的な応答と直接的な応答の収集のための指示書

Instructions

Read the following dialogue between the USER and the OPERATOR, please rephrase the USER's response written in red letters into two different types of speech, following the instructions below.

Type-1 (Direct) : a more direct response that expresses the same intention as the original response
Type-2 (Indirect) : a more indirect but natural response that expresses the same intention as the original response

"Indirect response" means, for example, a response to a Yes/No question that does not contain a "Yes" or "No", or a response that does not directly refer to the action you want the other person to do or your desire. If you have trouble rephrasing, click the "Hints" button. You can see the goals that "USER" must achieve in that interaction.

3. DIRECT コーパス

間接的な応答と直接的な応答の対からなる対話コーパスを構築する．データ作成コストの低減と，既存の対話システム研究との互換性の確保のために，本研究では既存の対話コーパスを拡張する形で応答対を収集する．具体的には MultiWoZ2.1（以降は MultiWoZ と表記） [3, 4] コーパスを基に，クラウドソーシングを用いて応答対を収集する．

本節ではまず応答対の収集方法と得られたデータ例について 3.1 節で説明する．また，収集したデータの品質評価を 3.2 節にて行う．構築した DIRECT コーパスの特徴や性質について，3.3 節では統計的な観点から分析する．

3.1 間接的な応答と直接的な応答の対の収集

MultiWoZ は 10,438 対話 からなるマルチドメインかつマルチターンのタスク指向対話コーパスであり，対話行為や対話状態など豊富なアノテーションが付属している．各対話はユーザとシステムの二者が交互に発話する形式であり，ユーザ発話は合計で 71,524 件存在する．

本研究ではクラウドソーシングサービスの Amazon Mechanical Turk[*2] を用いて，MultiWoZ を拡張する形で間接的な応答と直接的な応答の対を収集する．作業者にはまず表 1 の指示書と作業例を提示する．また，図 2 のように，各作業者の作業画面上には MultiWoZ から抜粋された対話履歴が表示される[*3]．表示された対話履歴に基づいて，各

[*2] https://www.mturk.com/
[*3] 元々の MultiWoZ データは システム（system）役とユーザ（user）

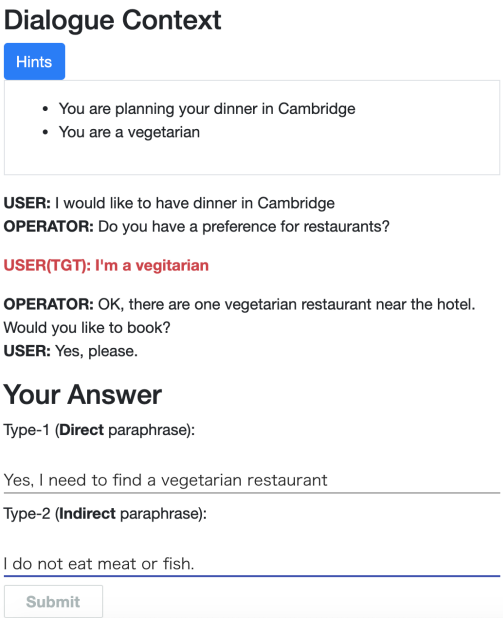

図 2 クラウドソーシングを用いた間接的な応答と直接的な応答の対の収集に用いる作業画面例

作業者は指定されたユーザ応答（図 2 中の赤字の応答）を間接的な応答と直接的な応答のそれぞれに言い換え，入力フォームに入力する．なお，対話履歴のみでは話者の意図を十分に理解できない場合に備え，「Hints」ボタンを押すことで MultiWoZ から抜粋したその対話の対話目標を表示できる機能を作業画面上に実装する．

1 件あたりの作業時間を平均 1 分と見積もり，平均報酬は 1 件あたり 0.12 米ドル（1 時間あたり 7.2 米ドル．2021 年 6 月 23 日現在のレートで日本円換算すると約 799 円）とした．最終的には 71,498 件[*4]の間接的な応答と直接的な応答の対を収集できた．得られたデータは，MultiWoZ と同じ設定で訓練データと評価データに分割した．

データ例

収集した間接的応答と直接的応答の例を表 2 に示す．上段の例におけるオリジナルのユーザ応答（MultiWoZ に元々収録されている応答）を見ると，ユーザは中間的な価格帯のレストランを希望していることがわかる．この応答を間接的に言い換えたものとして “I don’t want to overspend but remember its also vacation”（あまり使いすぎたくはないけれど，休暇でもあるということを忘れないで）という発話が得られた．この例では “its also vacation” というフレーズが直接的な応答における “not too cheap”（安すぎない）に対応している．下段の例では，間接的な応答の中に “Do you know of any in town?” というフレーズがある．対話履歴を考慮すれば，これは単に「何か街にあるもの」を知っているかどうかを問うための発話ではなく，直接的な応答における “Can you find me a guesthouse...?” のように「ゲストハウスを探してほしい」という要求を伝えるための発話であることが汲み取れる．

表 2 DIRECT コーパスのデータ例．“USER (間接的)” と “USER (直接的)” はそれぞれ本研究においてクラウドソーシングによって収集した間接的応答と直接的応答で，“USER (オリジナル)” は MultiWoZ から抽出したオリジナルの応答である．

話者	発話
	(中略．ユーザはレストランを探している)
SYSTEM	Would you like to pick a different type of food?
USER	Yes, what about British food please.
SYSTEM	What price range are you comfortable with?
USER (オリジナル)	Something in the moderate price range would be good.
USER (間接的)	I dont want to overspend but remember its also vacation.
USER (直接的)	Can you choose something that is not too expensive and not too cheap.
SYSTEM	Do you have a preference as to what area of town you dine in?

話者	発話
USER	I need a place to stay in the north
SYSTEM	OK im seeing alot of choices in hotels is there anything else You need in the hotel that would help narrow it down
USER (オリジナル)	I’d really like to stay in a guesthouse. I heard the ones in Cambridge are very nice.
USER (間接的)	I am thinking of staying in a guesthouse. Do you know of any in town?
USER (直接的)	Can you find me a guesthouse in Cambridge?
SYSTEM	How about the Acorn Guesthouse? It is rated 4 stars and is in the moderate price range.

3.2 品質管理

高品質な応答対を得るために，我々は事前にスクリーニングを実施し，作業者を選定した．また，クラウドソーシングを用いた品質評価も実施した．

作業者の選定

数単語の置き換えや入れ替えしか行っていないデータや，オリジナルの応答の言い換えにもなっていないようなデータの収集を防ぐため，本番の収集タスクに登用する作業者を事前に選定した．具体的には，2 回のパイロットタスクを実施した．なお，2 回目のタスクにおいては 1 回目のタスクにおいて作業者から受けた質問や得られたデータの品質などを基に指示書を部分的に修正し，本番のタスクにおいて使用したものと同一の指示書を使用した．本番タスクにおいては，間接的な応答と直接的な応答の間の単語レベ

役による二者間対話からなる．しかし本研究においては，人間同士の対話において自然に発生するような間接的な応答を収集するために，システム役の発話を “operator” による発話として表現し，人工的に生成されたものだと認識されないようにする．

*4 自然言語文として成り立っていない 26 件のデータを除外した

表 3　評価データの品質評価結果

Metric	Ratio [%]
Intention-accuracy (間接的)	95.0
Intention-accuracy (直接的)	99.7
Directness-accuracy (Exact)	81.4
Directness-accuracy (Relaxed)	89.4

ルの Jaccard 係数が 0.75 以上のデータを自動的に不採択にした．また，得られたデータの抜き打ち検査も実施した．最終的には，これらの手動評価と自動評価に合格した 655 人中の 536 人の作業者によってデータが作成された．

品質評価

応答対の収集後，最終的な成果物の品質評価も実施した．具体的には評価セットに含まれる 7,372 件分のユーザ応答に対し，得られた間接的な応答と直接的な応答の対の品質をクラウドソーシングを用いて評価した．作業者には評価対象となる応答対とその対話履歴を提示した．その際，応答対には “Response-A”, “Response-B” のラベルをランダムに割り当て，どちらが間接的（あるいは直接的）な応答として収集されたデータかわからないようにした．作業者にはまず，応答対が実際に MultiWoZ のオリジナルの応答と同じ意図を表現できているかどうかを “Yes”, “No” の二値で評価してもらった．次に，Response-A と Response-B のどちらがより直接的かについても評価してもらった．なお，判断に迷った場合のみ “no difference” を選ぶことを許容した．1 件あたりの作業にかかる平均時間を 30 秒と見積もり，平均報酬は 1 件あたり 0.06 米ドル（1 時間あたりでは 7.2 米ドル．2021 年 6 月 23 日現在のレートで日本円換算すると約 799 円）とした．各応答対に対して 5 人の作業者による評価を収集した．最終的な評価値はその多数決によって決定した．なお，評価タスクにおいては，収集タスクに参加した作業者が自分が作成したデータを自己評価することがないように作業者の割り当てを行った．

評価結果を表 3 に示す．Intention-accuracy は収集した応答のうち，オリジナルの応答と同じ意図を表現していると判断されたものの割合である．間接的な応答では 95.0%，直接的な応答では 99.7% と高い割合のデータがオリジナルと同じ意図を持っていることがわかる．間接的な応答の Intention-accuracy は直接的な応答の場合より 4.7% 低い．これは，ユーザの意図を間接的に表現しているために，人間の評価者にとっても解釈が困難な曖昧な表現が多少含まれているためであると考えられる．Directness-accuracy は，直接的な応答として収集された応答が実際に直接的だと判断された割合である．“Exact” は “no difference” と判断されたデータを許容しない設定，“Relaxed” は許容した設定での Directness-accuracy であり，それぞれ 81.4%, 89.4% と高い割合となった．DIRECT コーパスにおいては，これらの評価ラベルも提供する．

表 4　収集した間接的応答と直接的応答の対の統計情報

尺度	値 [単語]
語彙サイズ（間接的な応答）	6,273
語彙サイズ（直接的な応答）	4,664
平均文長（間接的な応答）	15.6
平均文長（直接的な応答）	12.4

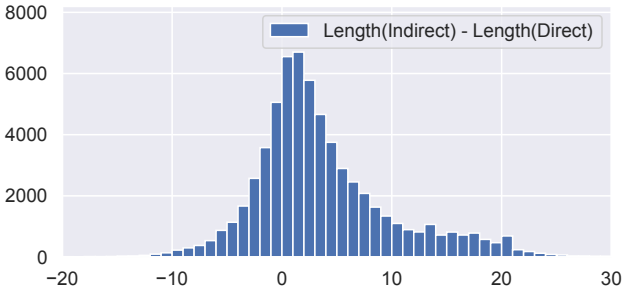

図 3　間接的な応答と直接的な応答それぞれの文長の分布

3.3　統計的分析

DIRECT コーパスにおける間接的な応答と直接的な応答の特徴を明らかにするために，まずトークン単位での統計的分析を行う．表 4 に単語レベルでの統計情報を示す．なお，単語分割には nltk[*5] ライブラリの ‘word_punct_tokenize()’ メソッドを用いた．また，大文字小文字は区別していない．まず，語彙サイズについては，間接的な応答の方が直接的な応答よりも 1.3 倍程度大きいことがわかる．これは，同じ意図を持った発話であっても，直接的に表現した場合よりも間接的に表現した場合の方がより多様な表現を取り得ることを示唆している．また，文長（応答に含まれる平均単語数）に関しては，間接的な応答では 15.6，直接的な応答では 12.4 で，間接的な応答の方が長い．平均文長について Wilcoxon の検定 [14] を実施したところ，0.1% 有意水準で有意であった．図 3 に，対応する間接的な応答と直接的な応答の文長の差のヒストグラムを示す．図を見ると，文長の差はやや正の側に偏りつつも，負の側にも多く分布していることが読み取れる．このことから，平均文長に有意差はあれど，発話をより直接的に言い換える際に単に文長を短くすることが必ずしも有効というわけではないことが示唆される．

次に，間接的な応答と直接的な応答において使用されるフレーズの違いを調査する．表 5 に，それぞれにおける単語 tri-gram の頻度上位 20 件を掲載する．表を見ると，直接的な応答の tri-gram には “book” や “find” などユーザがオペレータにして欲しいことを直接的に伝える動詞や，“the reference number” のように特定のオブジェクトを指すフレーズが頻繁に含まれていることがわかる（表中太字で表記）．一方，間接的な応答の tri-gram には，直接的な応答では頻出でない “is there any” や “I think that” など

[*5] https://www.nltk.org/

表 5　頻度上位 20 件の tri-gram

間接的な応答		直接的な応答	
trigram	freq	trigram	freq
i want to	2387	**find me a**	2617
i need to	2223	i want to	2442
would like to	1969	**can you find**	1971
i would like	1903	**please find me**	1924
is there any	1762	all i needed	1807
that would be	1588	thanks for the	1643
you help me	1584	in the centre	1628
thanks a lot	1407	i need to	1623
in the centre	1402	for the help	1539
i think that	1312	**you find me**	1531
i think i	1270	that's all i	1403
a place to	1246	can you get	1376
you have been	1242	give me the	1358
would be swell	1056	i need a	1206
such a great	1042	you get me	1200
i think you	1027	i would like	1145
you have done	1011	please give me	1097
a great help	981	get me a	1094
have been such	979	**book it for**	1086
been such a	979	**the reference number**	1060

のフレーズが含まれている．以上のことから，間接的な応答と直接的な応答では，それぞれ頻出するフレーズに違いがあることが読み取れる．

4. 間接的な応答の直接的な応答への言い換え

本節では間接的な応答を直接的に言い換えるタスクを設計し，実験を行う．本タスクの応用例として，タスク指向対話システムのための入力発話の前編集などが考えられる．例えば，ユーザが入力した間接的な発話を応答生成モデルに入力する際に事前に解釈しやすい直接的な発話に言い換えることで，応答生成の性能向上が期待される．

4.1 ベースラインモデルの構築

本タスクは入力された発話を意図を保持したままより直接的なスタイルに置き換えるスタイル変換タスクであると言える．そこで，文平易化 [15] やフォーマル性変換 [16] などのスタイル変換タスクにおいて高い性能を記録している BART [11] をこのタスクのベースラインとして用いる．ベースラインモデルのアーキテクチャを図 4 に示す．特殊トークンとして “<user>”, “<system>”, “<query>” を追加する．“<user>” トークン, “<system>” トークンをそれぞれ対話履歴中のユーザ発話の直前，システム発話の直前に付与することで，モデルがそれぞれの発話の話者を区別しやすいようにする．また，言い換え対象である間接的な応答の直前には “<query>” トークンを付与する．対話履歴中の発話と言い換え対象の応答を時系列順に連結し，BART のエンコーダに入力する．なお，入力データのトークン数

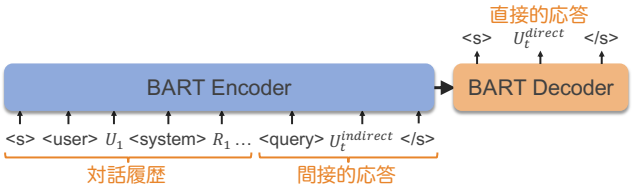

図 4　間接的な応答から直接的な応答への言い換えモデル

表 6　間接的な応答から直接的な応答への言い換えタスクの実験結果

Model	BLEU	Perplexity
Transformer w/ history	25.23	2.66
BART w/o history	32.51	**2.15**
BART w/ history	**33.77**	2.16

が 512 を超えた場合は末尾から 512 トークン目までのみを入力する．モデルのファインチューニングはクロスエントロピー損失を用いて行う．

実装には transformers [17] ライブラリを使用した．また，事前学習済みモデルとして “facebook/bart-base”[*6] を用いた．オプティマイザとして AdamW [18] を用い，学習率は 2e-5 [*7] とした．バッチサイズは GPU メモリ容量上の上限となる 8 を採用した[*8]．訓練データのうちランダムに 2,000 件をハイパーパラメータチューニング用の検証データとして抽出し，残りを訓練に用いた．30 エポックの学習を経て，最も検証データでの損失が小さいモデルを評価データでの評価に用いた．

また，対話履歴の影響を調査するために，対話履歴を入力しないモデル（BART w/o 対話履歴）も構築した．事前学習の効果を検証するために，事前学習されていない Transformer モデルを DIRECT コーパスのみを用いて訓練したモデルについても評価を行った．なお，Transformer モデルについてはエンコーダ・デコーダ共に層数や中間層の出力次元数などの各種ハイパーパラメータを BART と同一の値に設定した．

4.2 実験結果と考察

表 6 に各モデルの評価データにおける BLEU [19] と Perplexity を示す．BART を用いたモデルでは，対話履歴を考慮するモデルの方が BLEU スコアが高い．これは，間接的な応答は対話履歴を考慮しなければ正確に直接的な応答への言い換えができないことを示唆している．

BART と Transformer については BART の方が BLEU, Perplexity 共に大きく上回っていることがわかる．この結果から，他の多くの自然言語処理タスクと同様に，間接的な応答から直接的な応答への言い換えにおいても事前学習によって得られる知識が効果的に作用することがわかる．

[*6] https://huggingface.co/facebook/bart-base
[*7] 検証データにおいて最も損失が小さくなるように選んだ．
[*8] 使用した GPU は GeForce RTX 2080 Ti, 11GB Memory.

表 7 間接的な応答の直接的な応答への言い換えタスクにおける直接的な応答の生成例

USER	Hi, I'm looking for a place to eat some indian food.
SYSTEM	Do you have a price range in mind?
間接的な応答（入力）	I want it to be the best place in town.
BART	I want it to be an expensive place.
- w/o 対話履歴	Can you find me something expensive?
Transformer	Can you find me an attraction in town?
参照文	I was hoping for a much more expensive place if possible.
	(omitted)
SYSTEM	Yes, The Cambridge Belfry is a cheap hotel in the West.
USER	Do they have free internet?
SYSTEM	Yes, they have internet would you like me to book it for you?
間接的な応答（入力）	i dont think so on the booking. address and phone number though.
BART	dont need to book it. address and phone number for them though.
- w/o 対話履歴	address and phone number isnt needed.
Transformer	no...just give me the address and phone number.
参照文	address and phone number is all i need right now.

表 7 に生成された直接的な応答の例を示す．上段の例では，BART モデルは間接的な応答の “the best place” というフレーズを価格帯に関する表現だと解釈できていることがわかるが，一方で Transformer モデルはこの解釈に失敗している．下段の例は文脈の考慮が特に重要な例である．対話履歴を利用しない BART（BART w/o 対話履歴）では，参照文とは逆の意図を持つ文を生成してしまっている．一方，対話履歴を利用する 2 つのモデルは共に参照文と同じ意図の文を生成できている．

5. 対話応答生成への適用

対話システムにとって直接的なユーザ応答の方が間接的なユーザ応答に比べてより正しい応答を生成しやすいという仮定に基づき，入力発話を直接的に言い換えた発話を考慮する対話応答生成モデルの構築と評価を行う．具体的には，最先端の対話応答生成モデルである UBAR [20] に対し，入力発話に加えてそれをより直接的に言い換えたものも考慮するようにモデルを改良する．

5.1 直接的な発話への言い換えを考慮した対話応答生成器の構築

UBAR は事前学習済みのテキスト生成モデルである GPT-2 [21] を基にして構築された End-to-End 型のタスク指向対話応答生成モデルであり，MultiWoZ データにおける応答生成において高い性能を記録している．UBAR のモデルアーキテクチャを図 5 (a) に示す．UBAR では GPT-2 に対して対話履歴中の最初のユーザ発話を入力し，その入力を基に，ユーザが何を求めていて何が解決されていないか等，現状の対話状態を予測する．予測された対話状態を基にホテルやレストランの空き情報等が登録されたデータベースを検索し，条件に合致したレコードを GPT-2 に入力する（図中「DB」）．その後，次にどのような応答を生成するべきかを示す対話行為を推定し，最後にシステム応答を生成する．次のユーザ発話が入力された後も同様の処理を繰り返すことで，対話履歴を考慮しながら応答を生成していくようなモデルとなっている．

本実験では UBAR への入力ユーザ発話に対して，それと同じ意図を持つ直接的な応答を図 5 (b) のように連結して入力することで，入力発話を事前に直接的に言い換えた場合の応答生成性能を検証する．なお，予測対象の応答の直前のユーザ発話のみを言い換え対象としている．訓練時には直接的な応答として参照文（DIRECT コーパスに収録されている直接的な応答）を用いる．評価時には，第 4 節で構築した直接的な応答への言い換えモデルを用いて入力発話を直接的に言い換えたものを用いる．また，言い換えによる性能向上の上限を調査するために，参考として評価時に参照文を用いた場合の性能についても報告する．

訓練設定やパラメータについては全てのモデルについて Yang ら [20] での設定に準拠した．実装は著者らが公開しているスクリプト[*9]を用いた．MultiWoZ の訓練データと評価データへの分割方法も 4 節や Yang ら [20] の設定と同様である．

5.2 実験結果と考察

評価データにおける評価結果を表 8 に示す．なお，評価尺度は BLEU, INFORM, SUCCESS, COMBINED の 4 種類である．このうち INFORM, SUCCESS, COMBINED は MultiWoZ [3] 等のタスク指向対話データセットを用いた応答生成実験において代表的に利用されている評価尺度である．INFORM は対話全体を通して提示したエンティティ（ホテル名やレストラン名など）が正解と一致した割合であり，ユーザのニーズに合致した場所を案内できたかどうかを図る尺度である．SUCCESS は要求された属性値（住所や予約番号，列車番号など）を正確に出力できた割合であり，予約等の成功率を図る尺度である．COMBINED はこれら 3 つの尺度を統合したもので，$\text{COMBINED} = \text{BLEU} + 0.5 \cdot (\text{INFORM} + \text{SUCCESS})$ で計算される．

表より，全ての尺度において直接的に言い換えたユーザ応答を考慮したモデルが UBAR のスコアを上回っていることがわかる．特に，SUCCESS においては言い換えモデ

[*9] https://github.com/TonyNemo/UBAR-MultiWOZ

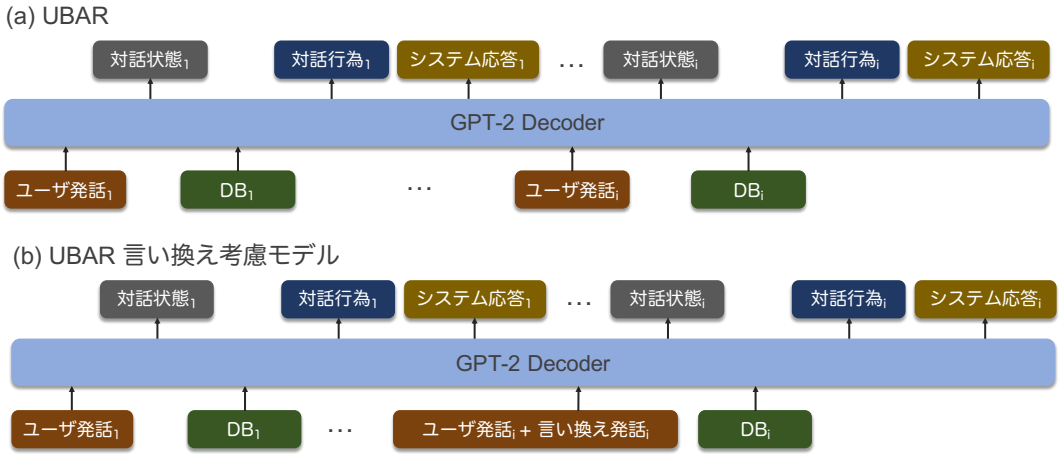

図 5 応答生成モデルのモデルアーキテクチャ

表 8 対話応答生成の評価結果

モデル	入力発話	BLEU	INFORM	SUCCESS	COMBINED
UBAR	オリジナルのみ	15.07	90.6	77.8	99.27
UBAR w/ 直接的な応答	オリジナル + 直接的な応答（生成）	**15.39**	91.1	78.8	100.34
UBAR w/ 直接的な応答	オリジナル + 直接的な応答（参照文）	15.27	**91.7**	**79.4**	**100.82**

ルによって生成された直接的な発話を用いた場合で 1.0 ポイント，参照文を用いた場合で 1.6 ポイント上昇している．よって，ユーザ発話の直接的な発話への言い換えは対話応答生成の性能向上に寄与することがわかった．また，生成文と参照文を比較した場合，BLEU 以外の全ての尺度で参照文の場合の方が性能が高いことが読み取れる．このことから，より正確な言い換えモデルを構築することで，さらなる応答生成性能の向上が見込める．

6. おわりに

本研究では 71,498 対の間接的な応答と直接的な応答の対を含む対話コーパスを構築した．また，間接的な応答を直接的な応答に言い換えるタスクを設計し，最先端の事前学習済み言語モデルの間接的な応答に対する処理能力を調査した．さらに，対話応答生成において直接的な応答への言い換えモデルを用いて入力発話をより直接的な発話に言い換えることで，応答生成の性能が向上することを確認した．今後はより性能の高い言い換えモデルの構築に取り組む．

本研究ではタスク指向対話を対象としてコーパスを構築したが，雑談対話においてはそのドメインの広さや発話の多様さから，より解釈の難しい間接的な応答が発生すると考えられる．雑談対話における間接的な応答に着目したコーパスの構築も将来的な課題として挙げられる．

謝辞 本研究は JSPS 科研費 JP18K11435 の助成を受けたものです。

参考文献

[1] Searle, J. R.: *Indirect speech acts*, p. 30–57, Cambridge University Press (1979).
[2] Li, Y., Su, H., Shen, X., Li, W., Cao, Z. and Niu, S.: DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Taipei, Taiwan, Asian Federation of Natural Language Processing, pp. 986–995 (2017).
[3] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O. and Gašić, M.: MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026 (2018).
[4] Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P. and Hakkani-Tur, D.: MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines, *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 422–428 (2020).
[5] Zhao, Y., Xu, C. and Wu, W.: Learning a Simple and Effective Model for Multi-turn Response Generation with Auxiliary Tasks.
[6] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J. and Dolan, B.: DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 270–278 (2020).
[7] Hosseini-Asl, E., McCann, B., Wu, C.-S., Yavuz, S. and Socher", R.: A Simple Language Model for Task-Oriented Dialogue, *Advances in Neural Information Processing Systems*, Vol. 33 (2020).
[8] Lin, Z., Madotto, A., Winata, G. I. and Fung, P.: MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*,

pp. 3391–3405 (2020).
[9] Pragst, L. and Ultes, S.: Changing the Level of Directness in Dialogue using Dialogue Vector Models and Recurrent Neural Networks, *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 11–19 (2018).
[10] Louis, A., Roth, D. and Radlinski, F.: "I'd rather just go to bed": Understanding Indirect Answers, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7411–7425 (2020).
[11] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880 (2020).
[12] Hou, Y., Liu, Y., Che, W. and Liu, T.: Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1234–1245 (2018).
[13] Gao, S., Zhang, Y., Ou, Z. and Yu, Z.: Paraphrase Augmented Task-Oriented Dialog Generation, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 639–649 (2020).
[14] Wilcoxon, F.: Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80–83 (1945).
[15] Martin, L., Fan, A., de la Clergerie, É., Bordes, A. and Sagot", B.: MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases, *arXiv preprint arXiv:2005.00352* (2021).
[16] Chawla, K. and Yang, D.: Semi-supervised Formality Style Transfer using Language Model Discriminator and Mutual Information Maximization, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2340–2354 (2020).
[17] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A.: Transformers: State-of-the-Art Natural Language Processing, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45 (2020).
[18] Loshchilov, I. and Hutter", F.: Decoupled Weight Decay Regularization, *International Conference on Learning Representations* (2019).
[19] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318 (2002).
[20] Yang, Y., Li, Y. and Quan, X.: UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2, *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence* (2021).
[21] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019).