

知識源との一対多関係を有する対話コーパスによる発話生成

金田龍平¹ 芳賀大地¹ 杉山弘晃² 酒造正樹¹ 前田英作¹

¹東京電機大学 システムデザイン工学部

²NTT コミュニケーション科学基礎研究所

{18aj045, 18aj110}@dendai.ac.jp

h.sugi@ieee.org {shuzo, maeda.e}@mail.dendai.ac.jp

概要

End-to-end 発話生成において、発話文脈に加えて外部知識を活用することにより、知識を反映した発話を生成できることが知られている。外部知識に基づく発話生成モデルに用いられる学習コーパスとして、Wizard-of-wikipedia のような、知識を参照しながらユーザに対話させ、その際の発話と知識の対を収集する枠組みが提案されている。しかし、ある文脈において利用可能な知識は一意とは限らず、実際に利用された知識以外にも利用可能な知識は存在する可能性がある。本研究では、旅行代理店における客と店員の対話を題材として対話データを収集し、基準対話データセットとした。基準対話データセットの一部に対し、発話文脈から利用可能な外部知識にリンクを付与することで、マルチラベル対話データセットを構築した。その結果付与されたリンクの総数は、基準対話データセット 1 に対し、マルチラベル対話データセットが 5.8 であった。このようなマルチラベル対話データを知識選択モデルの評価に用いたところ、知識選択モデルの適切な評価には発話と知識が一対多となるデータセットが良いことがわかった。また、発話生成モデルの生成に用いたところ、多様で適切な応答が可能になることが示唆された。

1 はじめに

対話システムにおける発話生成には、対話履歴などの文脈情報に加えて外部知識を利用する手法が提案されている [1, 2, 3]。発話生成部の学習は、対話履歴、知識情報、さらにそれらを利用して生成されるであろう後続発話からなるコーパスが必要になる。

そうした対話コーパスとして、児玉らによる映画推薦タスクにおける対話コーパスがある [4]。このコーパスでは、ある映画における監督名やキャスト

名などの構造化された知識に加え、あらすじやレビュー情報といった構造化されていない知識も外部知識源として利用している。そして、これらの外部知識を用いた発話生成を行い、20 ターン以上の約 2,500 件からなる対話データが収集されている。

児玉らの作成したコーパスでは、発話に付与された知識ラベルは、その発話を生成するのに使われたものとして知識候補の中から一意に定まる。すなわち、児玉らの作成したコーパスにおける発話と知識源の関係は、「先行発話と使用した知識」である。しかし、先行発話に対する応答が固定されていなければ、利用可能な知識は複数存在するはずであり、「先行発話と利用可能な知識」の関係も利用可能になると期待できる。したがって、このようなコーパスを用いて、知識を選択するモデルを評価することは適切なラベルで評価していないことになる。また、利用可能な知識が少なければ、生成できる応答文の多様性も低下してしまう。

そこで、2021 年に開催された対話ロボットコンペティション¹⁾のタスクである旅行代理店対話に着目し、外部知識に基づく対話コーパスを作成した。このコーパスは、クラウドソーシングによって作成した基準対話データセット、応答文生成に利用可能な知識に着目して作成されたマルチラベル対話データセットなどによって構成される。作成した旅行代理店対話コーパスの概要図と統計情報を、図 1 と表 1 にそれぞれ示す。マルチラベル対話データを用いた実験の結果、知識選択モデルを適切に評価するには発話と知識が一対多の関係となる評価データが良いことと、このようなデータセットを用いることで、多様かつ適切な応答が可能になることが示唆された。

1) <https://sites.google.com/view/crobotcompetition/home>

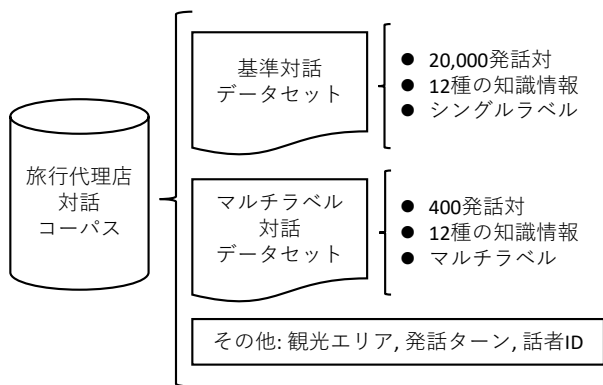


図1 旅行代理店対話コーパスの概略図

表1 旅行代理店対話コーパスの統計情報

対話数	2,971
発話対の数	16,285
単語数	581,763
異なり単語数	13,179

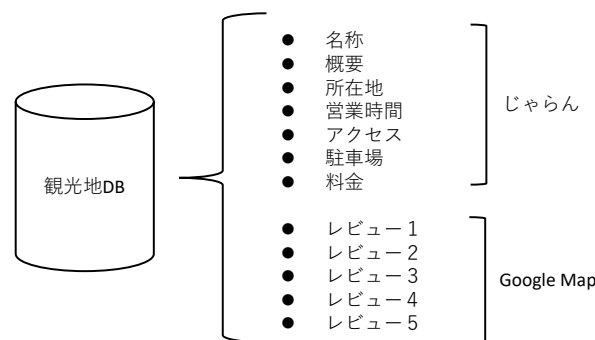


図2 知識情報の概要図

表2 知識情報の統計情報量

観光地	441
知識数	4,331
単語数	1,884,239
異なり単語数	19,347

2 関連研究

まず、日本語の対話コーパスに関する研究として、稲葉らの旅行代理店対話コーパス [5] や田中らのユーザの曖昧な要求に対して適切な応答をするための対話コーパス [6] などがある。稲葉らは、旅行代理店勤務経験者5名と7歳~72歳の男女55名による旅行代理店対話から、音声データやテキストデータなどを収集した。田中らは、タスク指向型対話において、ユーザが曖昧な要求をしてきたとしても対応できるように、旅行案内のドメインで対話データを収集した。

また、外部知識を用いた対話データとして Wizard-of-wikipedia [7] がある。これは、wikipediaを直接検索し、それによって得られた情報を用いて対話を行うことで外部知識と対話がセットになったデータセットを作成した。質問分類タスクに関する研究として、情報検索 (Information Retriever) の分野では、クエリと知識を入力とすることで分類問題を解く手法が提案されている [8]。この手法では、まず、知識候補をコーパス全体から BM25 [9] に従って数千程度に絞りこむ。そして、クエリと知識候補を入力としてクエリと知識候補の類似度を BERT に予測させることで、知識候補をリランキングし、最終的に回答生成モジュールに渡す知識を決定する。

3 旅行代理店対話コーパスの構築

3.1 基準対話データセット

クラウドソーシングを活用することで、東京と大阪の観光地441件を対象に、観光地に関する対話を合計で約3,000対話収集した。観光地には、事前に12種類の知識情報が与えられている。知識情報は、名称・概要・所在地・営業時間・アクセス・駐車場・料金・レビュー1~5といった12のカテゴリで構成される。名称・概要・所在地などの基礎情報はじゃらん²⁾からスクレイピングを行い、レビュー情報については Google Map API³⁾を用いて取得した。概要図を図2に示し、統計情報は表2に示す。クラウドワーカー (以下、ワーカー) は、店員の発話と客の発話をそれぞれ7発話作成する (以下、店発話、客発話)。このとき、店発話には上記で示した知識情報をなるべく用いて発話を作成してもらった。また、使用できる知識情報源は最大で2つとし、相槌などの知識情報を使用しない発話には、知識を使用しなかったという意味で none ラベルを付与した。具体的な作業手順と作業画面は、付録 A.1, 付録 A.2 にて示す。

これらの作業手順に従い、1つの観光地に対して、複数人のワーカーが作成することで合計2,971対話収集することが出来た。

2) <https://www.jalan.net/>

3) <https://developers.google.com/maps/documentation?hl=ja>

店発話	客発話	外部知識	生成文
東京旅行なら、まるごとっぽんがお勧めです。	それほどのようなところなのですか？	浅草にオープンした、日本全国の伝統や食文化・製品を紹介する施設。全国の隠れた名産品などを取り扱っている。日本の文化を見て、食べて、購入することができる注目のスポット。	全国の隠れた名産品を扱っているんです。
		お料理はどれも美味しい上に接客もほどよい距離感で居心地よし。外にはスカイツリー浅草寺や周辺をたくさん歩いてお腹を空かせてふと入ったお店です。定食のご飯の重さに驚きましたが気がついたら完食してしまい更にびっくりおつまみ豊富で悩めますよ。焼きだだちゃ豆の固さ最高でした！ごちそうさまでしたまた必ず行きますね！	定食のおつまみが充実しているお店です。
		平日のランチで利用しました。焼魚定食（太刀魚）1500円鰯たたき定食1400円煮魚（単品・金目鯛）800円黒烏龍茶500円4階にありました。窓の外には浅草寺とスカイツリーが並んで見られて、最高でした。店員さんの静かな話し方、接客も落ち着いていて、ゆっくりと食事を楽しむ事ができました。ご飯のお代わりも笑顔で対応してくれて、嬉しかったです。また利用したいと思います。	和食のランチです。定食は1500円、煮魚は800円とリーズナブルですよ。
		お料理がすべて美味しかったです。浅草でお手頃な和食屋さん店員さんも気取ってなくて程よい距離感で接してくれました。お料理に合うお酒も出してくれて大満足でした。	和食屋さんです。店員さんも程よく気取ってなくて、程よい距離感で接してくれるんですよ。
		お料理がとても美味しいです。また店員さんの対応も素晴らしいと思います。今ではお気に入りの店の一つです。	お料理がとても美味しく、店員さんも親切で評判がいいお店です。
		広くて綺麗。食事も美味しかったです。テラス席もあるので、もう少し涼しくなったらテラス席も良いかも。	テラス席がございます。外を眺めながら食事が楽しめる、有名なスポットです。

図3 外部知識を用いて生成した応答文の一例

3.2 マルチラベル対話データセット

応答文生成に利用できる知識は、発話と知識が「先行発話と利用可能な知識」という関係においては、複数存在すると考えられる。しかし、基準対話では、最大二つの知識を用いて応答文を作成している。したがって、ある発話に対して利用可能な知識は、最大二つまでとなってしまい、他の知識は利用できない知識ということになる。この利用できない知識は、本来「利用していない知識」であって「利用できない知識」ではない。そこで、400発話対（店発話+客発話）を用いてアノテーションを行った。アノテーションの対象としたデータは、基準対話データセットから400発話対を抽出し、それぞれに知識を付与した形へ整えたものである。ラベルは、ある発話対から次の応答文を生成する際に、付与された知識が利用可能かどうかを意味する。

マルチラベル対話データセットを作成する際に、単に整合性が取れているかだけでなく以下の点にも着目してアノテーションを実施した。

- 機械が対象の知識を用いて適切な応答文に変換できるか
- 客発話が相槌のような場合は知識におすすめポイントが入っているか
- 知識の内容がマイナスな内容ではないか

また、生成モデルに渡す知識情報が長くなると、モデルが知識情報を生成文に反映させることが難しくなると考え、概要とレビュー情報に関しては知識情

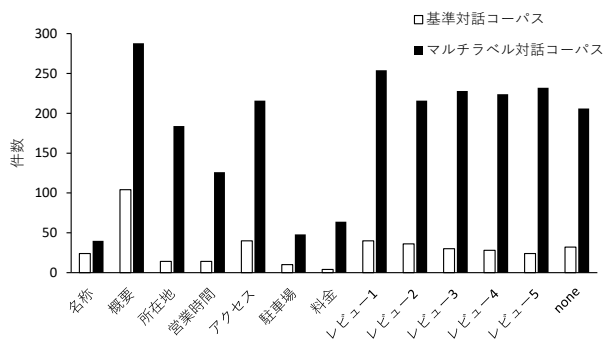


図4 基準対話データセットとマルチラベル対話データセットそれぞれにおいて、12種類の各知識源と関連付けられた発話例の総数。

報を句点で区切る処理を行った。この前処理を施したのものに対してもアノテーションをしたが、本稿では発話と知識情報の関係性に着目しているため、詳細は付録A.3にて述べる。アノテーションを実施した結果、利用可能な知識がどの程度増加したかを図4に示す。基準対話データセットの分布とマルチラベル対話データセットの分布を比較すると、多くの知識源が利用可能であることが分かる。すなわち、基準対話データセットでは、正解となる知識以外は利用できない知識となってしまうが、実際は利用できる知識が多く含まれていることを示している。

4 実験

4.1 実験概要

対話システムでは、外部知識を用いた生成を行う場合に、知識候補から最適な知識を選択する必要がある。

る。そこで、BERT を用いて利用可能な知識情報を選択し、Transformer ベースの大規模対話モデル（以下、hobbyist）で、知識情報を用いた応答生成を行う対話システムを構築した。BERT と hobbyist を学習する際に、基準対話データセットを用いた。また、hobbyist は、NTT コミュニケーションシステム基礎科学研究所が構築した [10]。

本実験では、知識選択モデルの評価にマルチラベル対話データセットを利用した場合に、基準対話データセットと比較してどのように変化するか確認した。また、先行発話に対する応答文を作成する際に、利用した知識から利用可能な知識として用いることのできる知識数を増やすことで、文脈として正しくかつ多様性のある生成が可能であるかも確認した。

4.2 実験設定

Laboro 社の提供している日本語版 BERT の事前学習済みモデル⁴⁾を用いた。BERT に用いる学習データは、基準対話データセットから以下の手順に従って作成した。

1. 対話データから発話対と知識情報を時系列に抽出（ただし、知識を使用していない発話対は除く）
2. 時刻 t の店発話と客発話に、12 ある知識情報のうち 1 つを結合
3. 店発話、客発話、知識情報の間にセパレーショントークン（[SEP]）を挿入
4. 正解ラベルは、時刻 $t+1$ の店発話作成に使用した知識情報であれば 0 を、そうでない知識情報であれば 1 を付与

また、評価時には上記手順によって作成された学習データを 8:2:2 = train:validation:test と分割したものと、マルチラベル対話データセットから上記手順に従って評価データを作成し、これを 1:1=validation:test となるように分割したのを用いた。基準対話データセットから作成した test データをシングル test とし、マルチラベル対話データセットから作成した test データをマルチ test とする。評価指標には正解率を用い、発話対に対して適切な知識を選択することが出来れば正解となる。

4.3 実験結果

まず、BERT を用いた知識選択の結果を示す。学習したモデルの評価結果はシングル test が 0.46、マルチ test が 0.90 であった。本実験では、ある発話対に対して適切な知識を選択することが出来れば正解となる。したがって、マルチラベルであるマルチ test に対する精度がシングルラベルであるシングル test と比較して高いのは妥当である。

次に、シングル test とマルチ test の両方を用いて hobbyist による発話生成を行った。生成文の一例を図 3 に示す。構築したマルチラベル対話データは人間が、利用可能であると判断したラベルが付与されている。図 3 にあるように、人が利用可能と判断したものであっても、生成モデルはすべて文脈として正しく、知識を反映した生成をしている結果となった。

4.4 考察

BERT の最適な知識選択において、正解率が 0.46 から 0.90 に上がったことから、シングル test で誤っていると判断された知識は、実際には利用できる知識であるということであり、既存手法における使用した知識という観点での最適な知識では過度に厳しい評価を行っていることが考えられる。また、実際に hobbyist が生成した発話文から、複数の利用可能な知識を用いて発話を複数生成し、その中から適切な発話をリランキングなどの処理によって最終的な発話を生成するような手法に、マルチラベル対話データセットのような複数の知識が可能であるデータセットを用いることで、多様かつ適切な発話が可能になることが考えられる。

5 おわりに

本稿では、発話対と知識源が一对一となる基準対話データセットを作成した。このデータセットを分析したところ、発話対と知識源の関係が、一对多が望ましいことが明らかになった。そこで、発話対と知識源が一对多となるマルチラベル対話データセットを作成し、これらを用いて実験を行った。実験では、BERT を用いた知識情報の選択と、hobbyist を用いた発話生成を行った。その結果、発話対に対する知識源が複数存在データセットを用いることで、多様で且つ文脈情報を考慮した発話生成が可能となることが示唆された。

4) <https://github.com/laboroai/Laboro-BERT-Japanese>

謝辞

本研究は JSPS 新学術研究 JP19H05693 の助成を受けた。

参考文献

- [1] Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. A neural network approach for Knowledge-Driven response generation. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 3370–3380, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.
- [2] Sixing Wu, Minghui Wang, Dawei Zhang, Yang Zhou, Ying Li, and Zhonghai Wu. Knowledge-aware dialogue generation via hierarchical infobox accessing and infobox-dialogue interaction graph network. In **Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence**, California, August 2021. International Joint Conferences on Artificial Intelligence Organization.
- [3] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. TopicKA: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence**, California, July 2020. International Joint Conferences on Artificial Intelligence Organization.
- [4] 児玉貴志, 田中リベカ, 黒橋禎夫. 外部知識に基づく発話生成に向けた日本語映画推薦対話データセットの構築. 言語処理学会第 27 回年次大会, 2021.
- [5] 稲葉通将, 東中竜一郎, 千葉祐弥, 駒谷和範, 宮尾祐介, 長井隆行. 多様な年代の話者による旅行代理店タスク対話コーパスの収集と分析. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 192–197, 2021.
- [6] 田中翔平, 吉野幸一郎, 須藤克仁, 中村哲. 曖昧な要求と気の利いた応答を含む対話コーパスの収集と分類. 言語処理学会第 27 回年次大会, 2021.
- [7] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In **International Conference on Learning Representations**, 2019.
- [8] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. **arXiv [cs.LG]**, January 2019.
- [9] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. **Foundations and Trends in Information Retrieval**, Vol. 3, No. 4, pp. 333–389, 2009.
- [10] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chat systems. September 2021.

A 付録

A.1 対話収集の教示文

対話データ収集のために、ワーカーに示した作業手順と、注意事項を図5に示す。

1. 作業について
 1. 1人で店員と客の二役をこなして対話文を作成。
 2. 「名称」～「料金」の情報を確認。
 3. 各7ターンの対話（店7発話、客7発話の合計14発話）。
 4. 対象の観光地をどのような「観点」に沿ってお勧めするか選択。
 5. 店員の発話は、可能な限り与えられた情報を用いて作成。
 6. 利用した知識は「利用知識」に記載し、最大で2つまで選択。
2. 注意事項
 1. 対話中、客は1回は質問とそれに対する回答を行う。
 2. 対話中、店員は1回は客の興味や経験を尋ね、より客に合ったお勧めをできるようにする。
 3. 店員の発話は、自身の知識や別途検索した結果ではなく、与えられた情報に基づいて発話を作成する。
 4. 概要などの与えられた情報は、あくまで参考にしてコピーしない。
 5. 発話数が7ターンになっているか確認する。
 6. 観点を選択しているか確認する。

図5 具体的な作業手順と注意事項

A.2 対話収集時の作業画面

対話コーパスを作成するにあたって、クラウドワーカーに依頼した作業ファイルを図6に示す。ワーカーは指示書に従って、該当のExcelファイルに対して、記入する形とした。

対象旅行施設 名称(N)	ユニバーサル・スタジオ・ジャパン (USJ)
概要(S)	オープン初年度に1100万以上のゲストが来場したこと話題を集めた、子供から大人まで楽しめるテーマパーク「ユニバーサル・スタジオ・ジャパン (USJ)」ハリウッド映画をテーマにしたアトラクションから、話題のキャラクターをテーマにしたものまでさまざまな楽しみ方ができるテーマパークで、その中でもウィザーディング・ワールド・オブ・ハリリー・ポッターは大人気！また、人気のアニメやゲームとのコラボレーションや、ハロウィンやクリスマスなど季節ごとに楽しめるイベント、アーティストによる音楽ライブなども定期的に開催し、若者だけでなく家族みんなで楽しめるテーマパークになっています。
レビュー-1(R1)	ユニバーサルすごく楽しかった！！ハリポッターのアトラクションは最新って感じで映像がリアルで、4Dみたいですが！がしかし乗り物酔いする人は注意が必要！気持ち悪くなつ
レビュー-2(R2)	関西のテーマパークと言ったら間違いないここですね。関東ではディズニーですが、USJの方がコンパクトにまとまっているので1日で回ることができ、満足感があります。また映
レビュー-3(R3)	1ヶ月以上前から予約して行きました。まさかのgo toトラベルがジャストで開始されて、人が少ない時に行く予定だったのにもみくちゃ状態に...(苦笑)それでも少ない方なんでしょう
レビュー-4(R4)	まず印象から初めていってきました。ホテルから見た感じより広く良かったです。内容はとても楽しかった。楽しめました。アトラクションハリリー・ポッターは夕方から2回も連続で
レビュー-5(R5)	イオンカードクリスマスナイト2020に当選したので、嫁と2人で電車で行きました②入場出来るのは14時から、19時から22時迄が貸切りとなります^_^ユニバーサル駅を降りて
所在地(L)	大阪府大阪市此花区桜島2丁目 1-3 3
営業日(D)	
営業時間(T)	営業時間：営業時間は予告なく変更する場合があります。来場前に再度ご確認ください。また、当日、パークオープン時間より早く入場を開始する可能性があります。予めご了承ください
アクセス(A)	(1)JRゆめ咲線ユニバーサルシティ駅から徒歩5分
駐車場(P)	駐車場案内 http://www.usj.co.jp/parkguide/access/carpark.html 駐車料金 http://www.usj.co.jp/parkguide/access/w_carpark_calendar.html
料金(M)	その他：料金はHPで確認ください。 http://www.usj.co.jp/ticket/index.html#
観点	(観点がその他の場合に自由記述)
ターン数-話者利用知識	発話
1.店員	
1.客	

図6 実際の作業画面

A.3 文区切りを施したマルチラベル対話データセット

使用した知識情報は、本文で示した通り12種類存在する。しかし、概要やレビュー情報にはテキスト長がとても長いものが含まれる。深層学習モデルには、扱えるトークン長があり、発話対に対して知識情報が極端に長いと知識の選択においても、発話生成においても、偏った学習を行ってしまうことが考えられる。したがって、概要とレビュー情報に関しては、句点で文を区切り、区切った後の一文で一つの知識情報として扱った対話データも作成した。これにより、概要とレビュー情報の知識情報が約1,200件から約3,200件に増加した。