対話エージェントの機能に着目した気の利いた応答を含むコーパスの収集

田中翔平 [1,3*]　吉野幸一郎 [2,1,3]　須藤克仁 [1,3]　中村哲 [1,3]
Shohei Tanaka[1,3]　Koichiro Yoshino[2,1,3]　Katsuhito Sudoh[1,3]　Satoshi Nakamura[1,3]

[1] 奈良先端科学技術大学院大学
[1] Nara Institute of Science and Technology
[2] 理化学研究所ロボティクスプロジェクト
[2] Robotics Project, RIKEN
[3] 理化学研究所革新知能統合研究センター
[3] AIP Center, RIKEN

Abstract: スマートスピーカーなどの音声アシスタントを動作させるタスク対話システムは，ユーザの要求に対して適切な機能を用いて応答を行う必要がある．従来のタスク対話システムは，ユーザの要求が明確であることを前提としているものが多く，曖昧な発話に対して気の利いた応答を生成することが難しい．そこで本研究では，ユーザの曖昧な発話，独話などに気の利いた行動を行うことができるシステムを構築するため，観光中のユーザ発話と，それに対するシステムからの気の利いた応答を含むコーパスを収集した．収集方法としては，API で定義される機能に紐づいたシステム側の応答に対し，その応答が気が利いているとみなせるような先行発話をクラウドワーカーに入力してもらった．はじめに，日英 2 言語に関して小規模収集を行い，2 つのコーパスを比較した．その結果，日本語コーパスの方が英語コーパスと比較して質が高いことが判明した．

1　はじめに

タスク対話システムとは，ユーザの要求に対してあらかじめ定義されたシステムの API などを用いて応答するシステムであり，スマートスピーカーやデジタルサイネージなどの形で実社会へと普及しつつある．しかしこれまで研究，実用化されたきたタスク対話システム [1, 2] は，ユーザ発話の中に明確に要求が含まれていることを前提としたものが多く，ユーザの要求が曖昧な場合に，適切な応答を生成することが難しい．一方で，気の利く人間のコンシェルジュやガイドなどは，例えば「ここの景色は綺麗だね」などの曖昧なユーザ発話に対して「写真を撮りましょうか？」などの「気の利いた応答」を行うことができる．

このように，要求が曖昧なユーザ発話に対してもユーザが望むであろう行動が可能なシステムを構築するため，本研究では，ユーザ発話と，それに対する対話エージェントの気の利いた応答を含むコーパスを収集した．ユーザとシステムの対話を想定したコーパスの収集方法としては，2 人の被験者をユーザ役とシステム役に割り当てて対話してもらう方法が一般的である [3, 4]．だが，ユーザ発話が曖昧である場合，そうした発話に対して常に気が効いた応答を返すことを，全てのシステム役発話者が行うことができるわけではない．既存のコーパス中からどの応答が気が利いているかを認定することも困難である．また，実際に気の利いた応答が認定できたとしても，システムが実行可能な行動はシステム自身に定義されている API などの機能に制約され，その応答を返すことが現実的ではない場合も多い．

そこで本研究では，システム側の応答をシステムが利用可能な API に基づいてあらかじめ定義し，その応答が気が利いているとみなせるような先行発話を，クラウドワーカーに入力してもらうことでコーパスを収集した．これは，API などに制約されるシステム応答に応じてこれらが有効な文脈を収集する方が，広範な文脈に対して現実的な気の利いた応答を収集できると考えたためである．

本研究ではこの仮説に従い気の利いた文脈･応答のペアを収集することができるか，日英 2 言語に関して合計 1,000 対話ほどの小規模収集を行った．収集したコーパス中のユーザ発話を「先行発話として不自然」「要求が明確である」「曖昧かつ先行発話として自然」の 3 つに分類した．本研究の目的は要求が曖昧なユーザ発話と，その発話に対応する気が効いたシステム応答の収集であるため，3 つ目に分類されるユーザ発話の割合が高いほど，コーパスの質が高いと言える．日英コー

*連絡先：奈良先端科学技術大学院大学
奈良県生駒市高山町８９１６番地の５
E-mail: tanaka.shohei.tj7@is.naist.jp

表 1: 対話エージェントの機能とカテゴリーのリスト

機能	カテゴリー	カテゴリー数
スポット検索	テーマパーク, 公園, スポーツ施設, 工房・体験施設, お土産・伝統工芸, 温泉・銭湯, 寺院, 神社, 城, 名所・旧跡, 自然・風景, 美術館, 博物館, 着物レンタル, 紅葉, 桜, 人力車	17
レストラン検索	寿司, カフェ, 子供向けレストラン, 餃子, 居酒屋, 懐石料理, 抹茶, かき氷, 和菓子, 洋菓子, パフェ, チョコレート, わらび餅, アフタヌーンティー, アイスクリーム, クレープ, 喫茶店, お好み焼き, 焼きそば, カレー, ラーメン, レストラン街, 精進料理, 食堂, そば, うどん, すき焼き, 天ぷら, 豆腐, 湯葉, とんかつ, うなぎ, ベジタリアン向けレストラン, 焼き鳥, ステーキ, おばんざい, 鍋, ハモ料理, 鴨料理, 焼き肉, イタリアン, フレンチ, ハンバーガー, ハンバーグ, しゃぶしゃぶ, たこ焼き, おでん, 中華, パン屋, 朝食, 価格帯が安いレストラン, 価格帯が普通なレストラン, 価格帯が高いレストラン	53
アプリ起動	カメラ, 写真, 天気, 音楽, 乗り換え, メッセージ, 電話, アラーム, ブラウザ, 地図	10

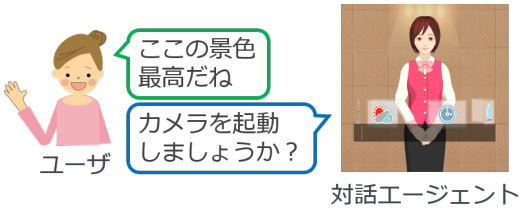

図 1: 気の利いた対話例

パスを比較した結果，英語コーパスと比較して日本語コーパスの質が高いことが判明した．これは日本語の場合と比較して，英語のクラウドソーシングにおける回答者の質を担保することが難しかったためだと考えられる．

2　コーパスの収集方法

本研究で収集するコーパスは，観光案内のドメインにおいてユーザとスマートフォンアプリケーション上の対話エージェントとの対話を想定したものである．対話は全て一問一答形式であり，ユーザは要求が曖昧な発話や独話を行い，対話エージェントはそのユーザ発話に対して気の利いた応答を返す．図 1 にユーザと対話エージェントの対話例を示す．ここでユーザの「ここの景色最高だね」という発話は必ずしも特定の機能に対する要求というわけではない．これに対して，対話エージェントが「カメラを起動しましょうか？」という気の利いた応答を返し，実際のカメラ機能を起動できるようにする．

図 1 のような対話を収集する方法として，クラウドソーシングなどを利用して，2 人のワーカーにユーザ役と対話エージェント役に分かれて対話してもらう (WoZ 対話) ことが考えられる．しかし，先に述べたように，意図の曖昧なユーザ発話に対して常に気の利いた応答を返すことは，人間にとっても難しいタスクであり，一般的な WoZ 対話では収集される対話の質を担保することが難しい．また，対話エージェントがスマートフォン上などで稼働するアプリケーションとして実運用されることを想定すると，その応答はアプリケーションが利用可能な API の機能に紐付いている必要がある．すなわち，対話エージェントにとって可能な気の利いた応答の範囲はユーザ発話の範囲と比較して限定的であり，あらかじめ定義した方がコーパスの質を担保しやすい．限定された対話エージェントの応答に対して，観光中のユーザの発話は多岐にわたる．そこで，広範なユーザの先行発話はワーカーに入力してもらうことで収集する．

本研究では対話エージェントの機能をスポット検索，レストラン検索，アプリ起動の 3 つに大きく分けて定義した．定義した機能のリストを表 1 に示す．各機能はそれぞれ細分化されたカテゴリーを持ち，コーパス中の対話エージェントの応答はこれらのカテゴリーに紐付いて生成される．スポット検索は特定のカテゴリーのスポットを検索する機能であり，「近隣の美術館を検索しましょうか？」といった応答としてユーザに提示される．レストラン検索は特定のカテゴリーのレストランを検索する機能であり，「近隣のかき氷を検索しましょうか？」といった応答としてユーザに提示される．アプリ起動は特定のアプリケーションを起動する機能であり，「カメラを起動しましょうか？」といった応答としてユーザに提示される．

定義した対話エージェントの応答に基づき，まず日本語コーパスをクラウドワークス[1]を用いて収集した．質の高いコーパスを収集するためには，対話の状況設定を明確にワーカーに伝達する必要がある．そうしたインストラクションの書き方を明らかにするため，異なる 3 種類のインストラクションを用いて，ユーザ発話をカテゴリー，インストラクションごとに 3 発話ずつ収集した．日本語コーパス収集におけるインストラクションおよび入力フォームの一例を図 2 に示す．図

[1]https://crowdworks.jp/

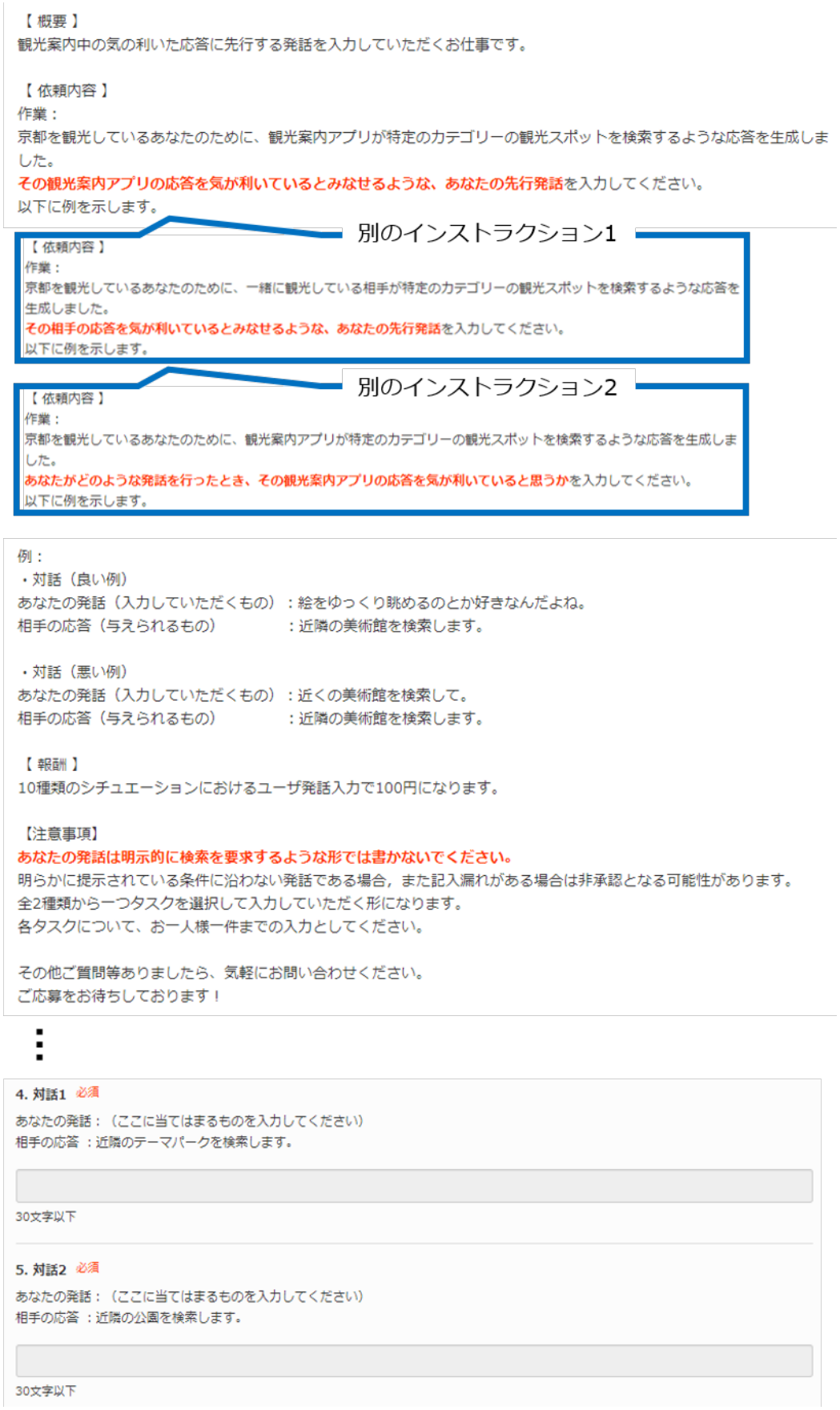

図 2: 日本語コーパス収集におけるインストラクションおよび入力フォーム．吹き出しはそれぞれ別のインストラクション．

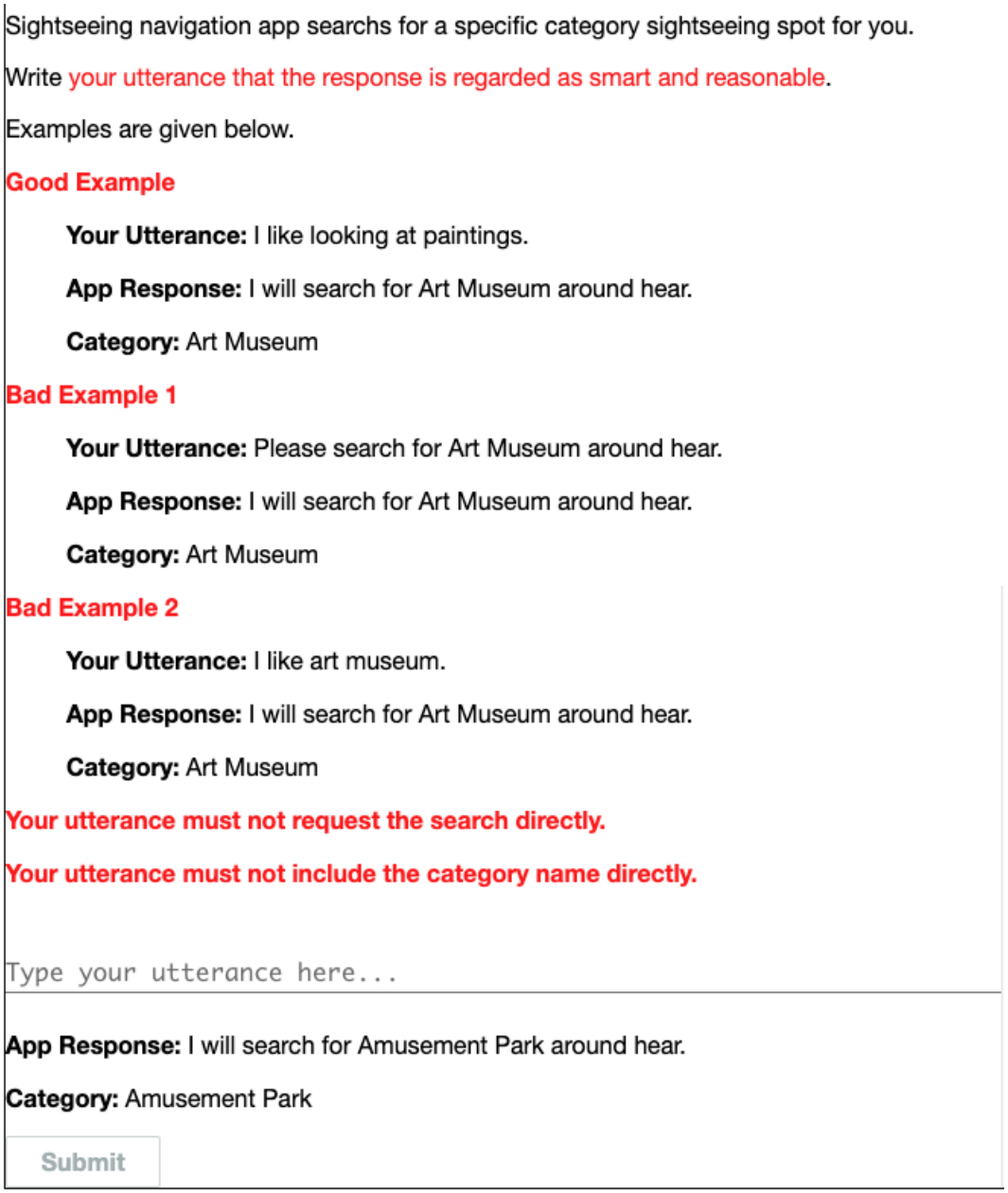

図 3: 英語コーパス収集におけるインストラクションおよび入力フォーム

表 2: 小規模コーパスの統計情報およびユーザ発話の人手評価．収集方法に不備があったため，英語のレストラン検索のカテゴリー数が日本語のものより 2 少なくなっている．括弧内の数字は割合を表す．

機能	言語	カテゴリー数	平均ユーザ発話長	先行発話として不自然 (%)	要求が明確である (%)	曖昧かつ先行発話として自然 (%)	データ数
スポット検索	日本語 1	17	16.14	0 (0%)	0 (0%)	51 (100%)	51
	日本語 2		16.43	0 (0%)	0 (0%)	51 (100%)	
	日本語 3		16.53	0 (0%)	7 (14%)	44 (86%)	
	英語		54.67	22 (43%)	16 (31%)	13 (25%)	
レストラン検索	日本語 1	53	15.09	2 (1%)	15 (9%)	142 (89%)	159
	日本語 2		15.72	2 (1%)	8 (5%)	149 (94%)	
	日本語 3		15.21	0 (0%)	17 (11%)	142 (89%)	
	英語	51	73.06	88 (56%)	44 (28%)	24 (15%)	153
アプリ起動	日本語 1	10	14.17	0 (0%)	10 (33%)	20 (67%)	30
	日本語 2		16.57	0 (0%)	1 (3%)	29 (97%)	
	日本語 3		15.03	0 (0%)	0 (0%)	30 (100%)	
	英語		48.77	16 (53%)	3 (10%)	11 (37%)	

表 3: コーパス中のユーザ発話例

ユーザ発話	システム応答	評価
思いっきり汗をかいちゃった． 最近，魚ばかり食べてるなあ． I don't know how to get to XXX.	近隣の温泉・銭湯を検索します． 近隣のステーキを検索します． I will search for a route to XXX.	曖昧かつ先行発話として自然
桜が見たいな． soba noodle is best for today ○○時に起こして．	近隣の桜を検索します． I will search for Soba Noodle around here. ○○時にアラームをセットします．	要求が明確である
i like looking at paintings 今日のランチは何を食べる． I love the view here.	I will search for Temple around here. 近隣のうなぎを検索します． I will launch Message app.	先行発話として不自然

中の吹き出しはそれぞれ別のインストラクションである．3 種類のインストラクションは，ユーザの対話相手が対話エージェントであるか，一緒に観光している相手であるかという状況設定や，ワーカーが入力する内容の説明の仕方が異なっている．本研究で収集するユーザ発話は要求が曖昧である必要があるため，良くない例として「近くの美術館を検索して.」という要求が明確な発話を挙げ，そのような発話は入力しないよう注記している．1 人のワーカーごとにそれぞれ異なる 10 カテゴリーに対してユーザ発話を入力してもらった．

次に，英語コーパスを Amazon Mechanical Turk[2] (MTurk) を用いて収集した．ここで用いたインストラクションは 1 種類のみであり，ユーザ発話の収集数は日本語コーパスと同様にカテゴリーごとに 3 発話である．英語コーパス収集におけるインストラクションおよび入力フォームの一例を図 3 に示す．インストラクションの内容は日本語のものとほぼ同様である．なお，1 発話あたりの単価は日本語と同じに設定した．

3 収集結果および評価

収集した各コーパスの統計情報を表 2 に示す．日本語と比較して英語の方が平均ユーザ発話長が長いが，これは日本語よりも英語の方が，同様の内容を表現するために必要な文字数が多いためだと考えられる．

収集したユーザ発話の質を明らかにするため，ユーザ発話を 3 種類に分類した．1 つ目の分類は「先行発話として不自然」であり，これは対話エージェントの応答が「カメラを起動します」なのにも関わらず，入力されたユーザ発話が「雨が降りそうかな」のような場合を指す．2 つ目の分類は「要求が明確である」であり，これは対話エージェントの応答が「カメラを起動します」なのに対して，入力されたユーザ発話が「カメラを起動して」のような場合を指す．3 つ目の分類は「曖昧かつ先行発話として自然」であり，これは対話エージェントの応答が「カメラを起動します」なのに対して，入力されたユーザ発話が「ここの景色最高だね」のような場合を指す．本研究の目的は要求が曖昧なユーザ発話の収集であるため，3 つ目の「曖昧かつ先行発話として自然」に分類されるユーザ発話の割合が高いほど，収集されたコーパスの質が高いと言える．

分類結果を表 2 に示す．日本語コーパス間の質の差は小さく，どの場合においても質が高いため，3 種類のインストラクションはどれも対話の状況設定を明確にワーカーに伝達できていることがわかる．だが，日本語コーパスの質と英語コーパスの質は大きく異なっている．日本語コーパスはどの場合においても「曖昧かつ先行発話として自然」に分類される発話が最も多い

[2]https://www.mturk.com/

のに対し，英語コーパスはどの場合においても「先行発話として不自然」に分類される発話が最も多くなってしまっている．英語コーパスに含まれる事例を調査したところ，インストラクションの例をそのままコピーしているものや，他の Web サイトなどからコピーしたと思われるものが多く見受けられた．このため，今後大規模収集を行う際は，クラウドワークスを用いて日本語コーパスを収集した方が質の高いコーパスを収集できると考えられる．また MTurk を用いて収集する場合は，禁忌肢問題[3]を設けるなど，質の低い回答者を自動で除外する仕組みが必要だと考えられる．収集されたコーパス中の発話例を表 3 に示す．5 番目の例はユーザ発話中に “soba noodle” という検索対象が含まれているため,「要求が明確である」に分類されている．8 番目の例は「今日のランチは何を食べる.」というユーザ発話中に，検索対象をうなぎに絞り込める情報が全く無いため,「先行発話として不自然」に分類されている．9 番目の例は “I love the view here.” というユーザ発話とメッセージアプリ起動との間の結びつきが弱いため,「先行発話として不自然」に分類されている．

4　おわりに

本論文では，対話エージェントの機能に着目し，ユーザ発話と対話エージェントの気の利いた応答を含むコーパスを収集する方法について述べた．具体的な収集方法として，API に基づいてあらかじめ定義された対話エージェントの応答に対し，その応答が気が利いているとみなせるような先行発話をクラウドワーカーに入力してもらった．日英 2 言語に関して小規模収集を行い，収集されたコーパスの質を比較したところ，日本語コーパスの方が英語コーパスよりも質が高いことが判明した．

今後は，まず日本語コーパスの大規模収集を行う．また，質の高い英語コーパスを収集するための仕組みについて検討していく．

謝辞

本研究にご協力いただいた理化学研究所革新知能統合研究センター観光情報解析チームの皆様に感謝いたします．

参考文献

[1] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1468–1478, 2018.

[2] Andrea Vanzo, Emanuele Bastianelli, and Oliver Lemon. Hierarchical multi-task natural language understanding for cross-domain conversational ai: Hermit nlu. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 254–263, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

[3] 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*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. 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)*, pp. 1951–1961, Hong Kong, China, November 2019. Association for Computational Linguistics.

[3]医師国家試験などで見られる，一定数誤答した場合即座に不合格となる問題.