常識推論に基づく気の利いた家庭内ロボットの行動選択

山﨑康之介 [1,2]　田中翔平 [1,2]　河野誠也 [2]　湯口彰重 [2,1] 吉野幸一郎 [2,1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所ガーディアンロボットプロジェクト
{yamasaki.konosuke.yi5,tanaka.shohei.tj7}@is.naist.jp
{seiya.kawano,akishige.yuguchi,koichiro.yoshino}@riken.jp

概要

日常的にユーザを支援するような家庭内対話ロボットでは、ユーザが明示的に行動を指示しない場合であっても、気の利いた行動を実行することが望ましい。例えば、ユーザが食事を食べ終えて「ごちそうさまでした」と発話した場合、潜在的には「テーブルに置かれたケチャップを片付ける」のような行動が期待される。こうした気の利いた行動を行うには、何らかの常識が必要であると考えられる。そこで本研究では、常識的な知識生成モデルを用いることで、発話からユーザが持つ潜在的な要求を推定し、気の利いた行動選択を行う。これにより行動選択と同時に推論過程を明確にし、提示することで気の利いた行動をユーザに受け入れやすいものにする。実験の結果、一部の発話では常識推論を用いることの有効性が示唆された。

1　はじめに

対話システム研究の発展によって、今後日常的にユーザの支援を行う対話ロボットの実現が期待される。ユーザの支援を行う対話ロボットは、ユーザの要求に応じて適切な行動を取る必要がある。しかしながら、人と共に生活し、日常的にユーザの支援を行う場合には、ユーザによる明示的な指示が無い場合であっても、発話や状況から気の利いた行動を推定し、実施することが期待される。

こうした対話ロボットの実現に向けて田中ら [1] は、非明示的な指示となり得るユーザ発話と周囲の状況をクラウドソーシングによって収集し、状況の再現動画を収録しデータセットを構築している。このデータでは、対話ロボットが一般的なリビングやキッチンにおいてユーザの家事を手伝うという状況を想定し、各エントリに対して 40 種類の行動候補から 1 種類の気の利いた行動が付与されている。田中らの深層学習に基づく行動選択実験の結果では、発話や周辺状況の画像を単に入力する場合の精度は低いが、周辺状況画像にある物体などをアノテーションした情報を含めると精度が大きく改善することが示された。つまり、気の利いた行動の推定には、ユーザが手にしている物体の検出といった周辺状況の細かな認識が重要である可能性が高い。

一方で、人間が気を利かせるような場合には、発話と周辺状況の詳細を把握するだけでなく、ユーザの直前の行動や何らかの背景知識を考慮した上で常識推論に基づき行動を決定することが考えられる。そこで、本研究では要求が非明示的なユーザ発話から常識推論によって背景知識を補完して気の利いた行動選択を行うシステムを構築する。

気の利いた行動はユーザにとっては飛躍した行動にも見えるため、行動に至った推論過程を同時に提示することが重要である。常識推論を利用することで行動決定の根拠を明らかにしつつ気の利いた行動の選択も実現できる。本研究では常識推論モデルを導入し、発話に関連する将来のユーザの行動などのイベントを推論する。常識推論には COMET [2] を再帰的に用い、この推論結果と行動候補の類似度を計算し、行動を決定する。例えば、ユーザの発話として「お腹がすいた」という発話があった場合、「“お腹がすいた”と言う-(その後)-> 食べ物を食べる-(その前)-> 食べ物を取る」という常識推論が展開できた場合を考える。この時、常識推論の結果である「食べ物を取る」と行動候補である「バナナを取る」は類似しており、何らかの類似度計算によって「バナナを取る」という行動が「お腹がすいた」という発話に対応する気の利いた行動になるという推論ができる。

気の利いたロボットの行動選択データセット [1] を用いて実験を行った結果、ベースライン手法の性能を上回ることはできなかったが、一部の発話においては、常識推論によって背景知識を用いながら明確なプロセスで推論を行う有効性が確認された。

また、現在のデータセットの改善の余地が示唆された。

2 常識推論に基づく行動選択タスク

本研究では、田中ら [1] が収集した家庭内におけるロボットの気の利いた行動選択データセットを用いる。データセットには、400 の発話とその周囲状況が収録されており、発話それぞれに対し、40 の行動カテゴリ（付録 表 3）から 1 つの正解行動が付与されている。例えば、ユーザが食事を終えて「ごちそうさま」と発話したという状況には「ケチャップを片付ける」という行動が正解として与えられている。これは、「ごちそうさま」という発話自体には要求や行動が明示的に表現されていないが、潜在的には、ご飯を食べ終えてテーブルの上のケチャップが不要になったのでロボットが気を利かせて片付けて欲しい、という要求が暗に存在すると解釈するものである。この例に対して常識推論を当てはめると、「“ごちそうさま”と言う」後に「食事に使ったものを片付ける」という常識推論ができれば、システムは行動を正しく選択できるということになる。本研究では発話からの推論可能性を模索するため、ユーザの発話のみを入力とした行動選択を行う。

3 推論システム

本研究では、発話から常識推論を行うために常識的知識に基づく推論を生成するモデルである COMET [2] を用いる。ここでは、COMET の利用方法と、推論システムの詳細について述べる。

3.1 COMET を用いた常識推論と前処理

COMET は、常識的な知識を表現した知識グラフ（Knowledge Graph; KG）である ATOMIC2020 [2] を用いて学習された、常識的知識の生成モデルである。ATOMIC2020 は、CONCEPTNET[1)] [3] のような既存の KG に比べて多様な知識を含んでおり、特に、イベントに関する知識が含まれている。例えば、「PersonX eats dinner」というイベントと、その後に起きるイベントを予測するための関係 *isBefore* を入力すると、「PersonX goes to bed」などの対応する関係を持つであろうイベントを出力する。つまり、発話に関するイベントと関係を入力することで、ユーザが発話後に取りそうな行動などを出力することができる。さらに、本研究では COMET の出力を新たな入力とすることで COMET を再帰的に用いる。この再帰の数を推論の深さと呼ぶ。

1) https://conceptnet.io/

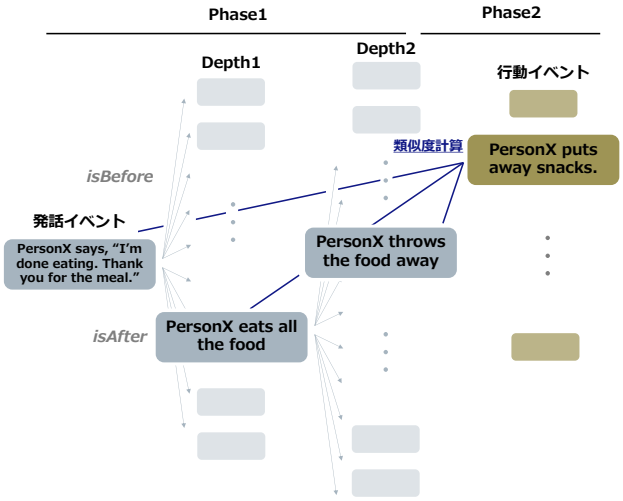

図 1 関係として [*isAfter*, *isBefore*] の 2 種類を使用した場合の推論システムの概要図。COMET は 1 種類の関係につき 5 つの生成を行い、最大の推論の深さは 2 である。Depth1 では、発話イベントを入力とした 1 回目の生成、Depth2 では、Depth1 の出力を用いて 2 回目の生成を行う。Phase2 では、発話イベントを含む全ての生成結果と行動候補の類似度に基づいて行動選択を行う。

COMET を気の利いた行動選択データセットに適用するため、次の 4 つの前処理を行なう。

- **翻訳**：COMET の入出力は英語である必要があるため、気の利いた行動選択データセット中の発話と行動を英語に翻訳する。
- **「PersonX says, “発話”」に変換**：COMET の入力はイベントの形式であることが望ましいため、ATOMIC2020 で人物を表す「PersonX」を主語に用いたイベント形式に変換する。
- **「PersonX 行動」に変換**：データセット中の翻訳後の行動は、例えば「put away ketchup」のように主語がないため、「Person X puts away ketchup」のように「PersonX」を主語に用いたイベント形式へ変換する。なお、「持ってくる (bring)」はユーザ側の行動として「get」に置換する。
- **関係に *xWant* を用いた場合の出力に「PersonX wants」を前置き**：関係に *xWant* を用いた場合、「to go to bed」など to 以降が出力されることが多いため、「PersonX wants」を前置きし、イベント形式へ変換する。

表 1　実験結果

	最大深さ 1			最大深さ 2			最大深さ 3		
関係	Acc.(%)	Top5-Acc.(%)	macro-F1	Acc.(%)	Top5-Acc.(%)	macro-F1	Acc.(%)	Top5-Acc.(%)	macro-F1
isBefore	**35.25**	59.75	**0.3376**	30.25	60.25	0.2906	19.75	57.25	0.1890
isAfter, *isBefore*	34.00	60.00	0.3136	28.25	**62.00**	0.2684	17.50	55.75	0.1720
xWant, *Causes*, *HasSubEvent*, *HinderedBy*, *isAfter*, *isBefore*, *xEffect*, *xReason*	33.50	59.00	0.3291	26.25	59.75	0.2517	12.25	53.00	0.1209
ベースライン（発話イベント）	39.50	63.00	0.3778						

3.2　推論システムの詳細

常識推論に基づく気の利いた行動選択は、以下の 2 段階の処理（Phase1, 2）によって実現した。また、図 1 に、推論システムの動作過程の具体例を示す。

Phase1　発話イベントを始まりとして、COMET を再帰的に用い、背景知識として次にとりうる行動などのイベントを複数生成する。この際、最大の推論の深さや用いる関係はあらかじめ決定しておく。深さが t のとき、COMET の出力 $Output_t^{i_t}$ は次のように表すことができる。

$$Output_t^{i_t} = COMET(Output_{t-1}^{i_{t-1}}, Relation_j) \quad (1)$$

ここで、用いる関係の種類が J 個、各深さで出力される推論の数が I_t 個、最大の深さが T であり、$1 \leq i_t \leq I_t, 1 \leq j \leq J, 1 \leq t \leq T$ となる。なお、$Output_0$ は発話イベントである。図 1 の例では、t=1 の時、$Output_1$ は 10 個生成される。

Phase2　Phase1 で得られた全ての出力と気の利いた行動候補との類似度を計算し、最も高い類似度を持つ行動を予測とする。出力と 40 個の各行動候補との類似度は SentenceTransformers[2] を用いて計算する。SentenceTransformers は Sentence-BERT [4] を基として、文同士の意味類似度を計算することができる。ここで、最も類似度の高い推論結果と行動候補のペアを求め、これを発話に対して選択された行動とする。行動イベントを $Action_k$ $(1 \leq k \leq 40)$、意味類似度の計算を $Similarity$ とすると、選択される行動 $Action$ と対応する推論過程は次のように求められる。

$$\hat{i_t}, \hat{t}, \hat{k} = \underset{i_t,t,k}{\arg\max}\, Similarity(Output_t^{i_t}, Action_k) \quad (2)$$

$$Aciton = Action_{\hat{k}} \quad (3)$$

これにより、$Output_{\hat{t}}^{\hat{i_t}}$ から遡ることで、発話イベントからの一連の推論過程を明示することができる。

2）https://www.sbert.net/index.html

4　実験

実験では提案システムを用い、データセットに含まれる 400 の発話それぞれに対して、40 の気の利いた行動候補から常識推論に基づく行動選択を行った。実験はいくつかの条件に分けて行い、3 つの評価指標で比較した。

4.1　実験設定

推論システムで利用する COMET は BART [5] ベースの demo モデル[3] を利用する。用いる関係は [*isBefore*]、[*isAfter*, *isBefore*]、[*xWant*, *Causes*, *HasSubEvent*, *HinderedBy*, *isAfter*, *isBefore*, *xEffect*, *xReason*] の 3 つの組み合わせを採用した。最大深さは 1 から 3 に設定し、評価指標には Accuracy、予測の上位 5 クラスまでに正解が含まれるかを反映した Top5-Accuracy、各カテゴリの F1 Score の平均（macro-F1）を用いた。SentenceTransformers は高速な計算が可能な all-MiniLM-L6-v2 モデル[4] 利用した。また、ベースライン手法として、COMET を用いずに発話イベントのまま、行動候補と類似度を計算して選択する手法を用いた。

4.2　結果

実験結果を表 1 に示す。どの指標においても最も良い性能であったのは、ベースライン手法であった。提案手法に注目すると、Accuracy、macro-F1 は [*isBefore*] のみを用い、最大深さを 1 にしたものが最も良く、Top5-Accuracy は [*isAfter*, *isBefore*] で最大深さ 2 が最大であった。Accuracy、macro-F1 は、用いる関係によっても違いが確認されたが、最大深さが大きくなるにつれての悪化が顕著であった。Top5-Accuracy の良さはいずれの関係でも、最大深さ 2>1>3 という順であり、Accuracy、macro-F1 の傾向とは異なった。いずれにしても最大深さを 3 に

3）https://github.com/allenai/comet-atomic-2020
4）https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

表 2　分析結果。C-Sim は正解の場合、W-Sim は不正解の場合の最大類似度の平均。W-Coverage はベースラインでの不正解の発話のうち、提案システムで正解した発話の割合。

	最大深さ 1			最大深さ 2			最大深さ 3		
関係	C-Sim	W-Sim	W-Coverage(%)	C-Sim	W-Sim	W-Coverage(%)	C-Sim	W-Sim	W-Coverage(%)
isBefore	0.8515	0.7582	11.57	0.8731	0.8265	10.74	0.8851	0.8803	7.02
isAfter, *isBefore*	0.8651	0.7811	11.98	0.8960	0.8516	11.16	0.9299	0.9054	5.37
xWant, *Causes*, *HasSubEvent*, *HinderedBy*, *isAfter*, *isBefore*, *xEffect*, *xReason*	0.8890	0.8119	12.40	0.9345	0.8911	10.33	0.9720	0.9633	4.96
ベースライン（発話イベント）	0.6999	0.5607	-						

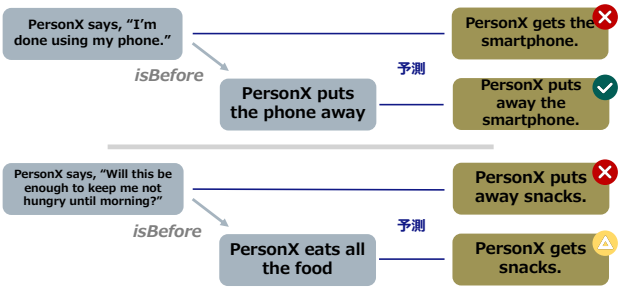

図 2　行動選択の実際の例。上の例は提案システムのみ正解したケースである。下の例は「Personx gets bananas.」が正解行動であるため不正解であったが妥当な推論であったケースである。

設定することは望ましくないが、最大深さ 2 では Accuracy 向上の見込みが示された。

5　分析

実験の結果、ベースライン手法は 3 つの評価指標のもとで最良の結果であった。ここではベースライン手法と提案システムの差異について評価指標以外の観点から分析を行う。

提案システムでは最大類似度のペアに含まれる気の利いた行動が選択される。各発話に対する最大類似度について、予測が正解であった場合と不正解であった場合の平均を表 2 に示す。提案システムは正解、不正解どちらの場合でもベースラインよりも類似度が高く、提案システムは意味的な類似度の上では実際の気の利いた行動とより近いイベントを推論できている。また、ベースラインを含むどの設定においても正解の場合の方が類似度が高い。つまり、気の利いた行動との類似度が高いイベントが存在する時に正解となる傾向である。提案システムでは不正解の場合にも比較的類似度が高いため、現在のデータセットでは不正解であるが、実際には妥当な行動が選択されている可能性がある。一方で COMET が発話と破綻した関係にあるイベントを生成している可能性もあり、今後の解析が必要である。次に、ベースラインで不正解であった発話のうち、提案システムで正解であった割合についても表 2 に示す。最も高い場合は 12.40% であり、ベースラインでは不正解となる発話についても提案手法では正解できるケースが一定数存在することを確認できた。

図 2 は実際の推論の例である。提案システムのみ正解できている例において、ベースライン手法では「phone」という単語に注目した類似度から行動を選択していると考えられるが、提案手法では常識推論を用いることで次の行動を推論し、気の利いた行動を選択できている。不正解の例では、「Personx gets bananas.」が正解であるため、予測した行動は不正解となっているが、尤もらしい常識推論によって気の利いた行動を選択できているように見受けられる。

類似度の傾向や実際の例からベースライン手法と提案システムを分析、再評価することにより、常識推論を用いて行動選択を行う提案システムの有効性や、利用したデータセットに複数の気の利いた行動候補を付与するなどの改善の余地が示唆された。

6　終わりに

本稿では、指示が明示的でない発話に対してシステムが適切な行動選択を行うことが期待される場面において、常識推論によってその曖昧性を補完し行動選択を行う枠組みを構築した。実験の結果、精度の課題は残るものの、一部の例では常識推論によって気の利いた行動に結びつく推論を実現できていることが確認できた。今後は、COMET のような単純な常識推論だけでなく、仮説推論に基づく手法 [6] や典型的な一連の行動を推論する手法 [7] にも着目し、発話の曖昧性の解消を試みる。また、データセットに付与された正解行動以外にも妥当な行動がある可能性や、あるいはデータセットに付与された気の利いた行動の難易度なども考慮しつつ、データセットについての議論も行う。

謝辞

本研究は JSPS 科研費 22H03654 の助成を受けた。

参考文献

[1] 田中翔平, 湯口彰重, 河野誠也, 中村哲, 吉野幸一郎. 気の利いた家庭内ロボット開発のための曖昧なユーザ要求と周囲の状況の収集. 情報処理学会研究報告 Vol. 2022-NL-253, No. 5, pp. 1-7, 2022.

[2] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 6384–6392, 2021.

[3] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In **Proceedings of the AAAI conference on artificial intelligence**, 2017.

[4] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[6] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning. In **International Conference on Learning Representations**, 2020.

[7] Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. proScript: Partially ordered scripts generation. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 2138–2149, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

付録

表 3　行動カテゴリ一覧

■持ってくる

バナナを持ってくる, 充電ケーブルを持ってくる, コップを持ってくる, ケチャップを持ってくる, 宅配便を持ってくる, ペットボトルを持ってくる, リモコンを持ってくる, スマホを持ってくる, お菓子を持ってくる, ティッシュ箱を持ってくる, 缶切りを持ってくる, クッキングシートを持ってくる, グラスを持ってくる, おろし器を持ってくる, キッチンペーパーを持ってくる, レモンを持ってくる, オリーブオイルを持ってくる, じゃがいもを持ってくる, サランラップを持ってくる, 水筒を持ってくる

■片付ける

充電ケーブルを片付ける, コップを片付ける, ケチャップを片付ける, ミニカーを片付ける, ペットボトルを片付ける, リモコンを片付ける, スマホを片付ける, お菓子を片付ける, ティッシュ箱を片付ける, 缶切りを棚にしまう, クッキングシートを棚にしまう, グラスを棚にしまう, おろし器を棚にしまう, キッチンペーパーを棚にしまう, サランラップを棚にしまう, 水筒を棚にしまう

■その他

ゴミをゴミ箱に入れる, ペットボトルを冷蔵庫にしまう, タッパーをレンジに入れる, タッパーを冷蔵庫にしまう