# 対話に基づく常識知識グラフの構築と対話応答生成に対する適用

井手竜也 [1]　榮田亮真 [1]　河原大輔 [1]
山崎天 [2]　李聖哲 [2]　新里顕大 [2]　佐藤敏紀 [2]
[1] 早稲田大学理工学術院　[2] LINE 株式会社
{t-ide@toki.,s.ryoma6317@akane.,dkw@}waseda.jp
{takato.yamazaki,shengzhe.li,kenta.shinzato,toshinori.sato}@linecorp.com

## 概要

コンピュータに常識を与えるため，多くの常識知識グラフが提案されているが，文脈を考慮したものは少ない．本論文では対話における文脈に注目し，常識に基づく対話応答生成に向けた対話常識グラフを提案する．カテゴリと時系列，対象といった次元を考慮し，提案するグラフを日本語で構築した．構築したグラフをもとに対話応答生成を行い，心情に関する推論を明示的に与えることによって生成される応答の特徴が変化することを示した．日本語で構築した対話常識グラフは公開予定である．

## 1　はじめに

人間は対話において，状況や相手が思うことを暗黙的に推論 [1] する．いわば人間は，対話の常識[1)] をもつと言える．それに対して，ニューラル雑談対話システムには対話の常識が足りず，普遍的な応答ばかりを生成するという指摘 [2] がある．コンピュータに常識を与える研究は盛んになされており，その一環として常識知識グラフ [3] が構築されている．

常識に関する推論は状況や文脈によって異なるが，既存の常識知識グラフは文脈を無視 [4] している．したがって，システムが文脈に依存した常識を推論するための，文脈を考慮した常識知識グラフが必要となる．英語にはいくつか存在 [5, 6, 7] しているが，日本語にはない．

本論文では，より人間らしい対話システムの実現に向けて，対話の常識を集めた対話常識グラフを提案する．カテゴリと時系列，対象といった次元に注目し，推論すべき関係を定義する．また大規模言語モデルを用いた対話応答生成に対して，対話常識グラフを適用する手法も提案する．

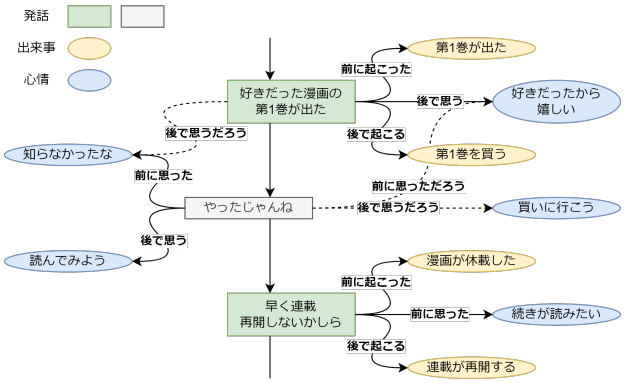



**図 1**　発話ごとにタグ付けする推論の例

提案する対話常識グラフを，日本語で構築[2)] した．Twitter API を用いて対話を収集し，Yahoo!クラウドソーシングを用いて発話ごとに推論を付与した．対話常識グラフの分析では，対話の常識における推論の傾向を明らかにした．さらに，構築したグラフを用いて対話応答生成を行った．HyperCLOVA JP [8] の In-Context Learning [9] において，心情の推論をプロンプトとして明示的に与えた．推論を与えることによって，生成される応答の特徴が変化することを示した．

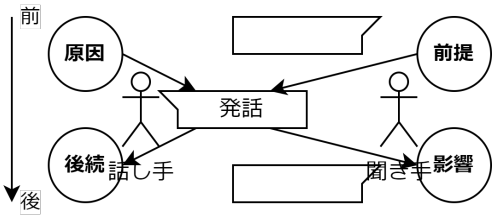
## 2　関連研究

### 2.1　常識知識グラフ

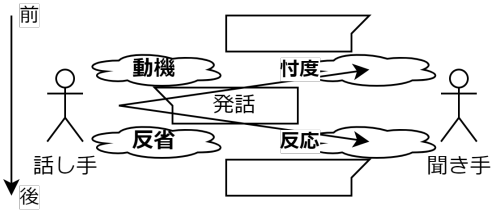
イベントに関する常識を集めた常識知識グラフに ATOMIC [3] があるが，それぞれのイベントは文脈を伴わない．それに対して PARA-COMET [4] は，物語における一連の文に対して推論を付与する．

1)　常識には Wikipedia に記載されるような知識に該当するものと，社会的相互作用にまつわるものがある．本研究では対話に関して，後者の常識を対象とする．

2)　日本語で構築した対話常識グラフは，https://github.com/nlp-waseda/dcsg-ja にて公開予定である．




(a) 出来事




(b) 心情

**図 2**　対話常識グラフで推論すべき関係

GLUCOSE [5] は物語の文に関する推論，CIDER [6] は対話における発話同士の推論を扱う．CICERO [7] も対話に関する常識知識グラフだが，テキストを超えた範囲の推論まで考慮する．

ATOMIC の関係は，対象がイベントの当事者かその他かを X と Others によって区別する．一方で CICERO の次元は，出来事か心情かと時系列の前後を区別しているが，推論の対象となる話者を区別しない．本研究では，さらにどの話者に向けた推論かという次元に注目する．

## 2.2　大規模言語モデルと対話応答生成

In-Context Learning [9] では，解くべきタスクの例をいくつかプロンプトとして大規模言語モデルに与え，そのタスクを解かせる．In-Context Learning を用いた対話応答生成 [10] の研究もなされている．特定の状況に基づく応答を生成する試み [11] や，共感的な応答を生成する試み [12] がある．

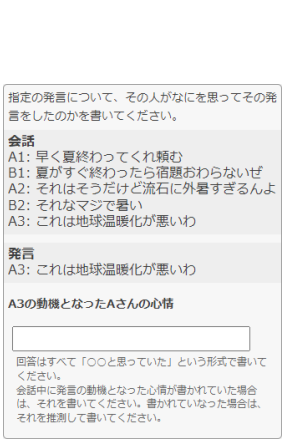
大規模言語モデルの生成は，与えられるプロンプトに大きく影響を受けること [13] が知られている．タスクを解くにあたって，その過程をあえてプロンプトに列挙する Chain of Thought Prompting [14] も提案されている．

# 3　対話常識グラフの構築

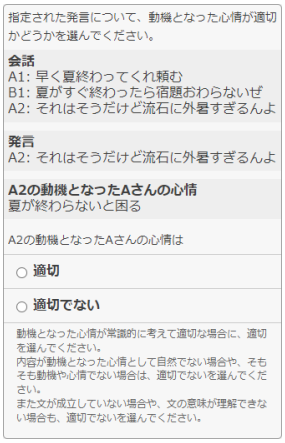
本研究では，対話の常識に特化したグラフを構築する．テキストに書かれているものから暗黙的なものまで，発話ごとに推論を付与する．発話ごとの推論は，より人間らしい対話応答生成に役立つと考える．

**表 1**　推論の統計

| 関係 | 数 | 平均数 | 平均文字数 |
|---|---|---|---|
| 原因 | 3,060 | 1.44 | 6.24 |
| 前提 | 2,728 | 1.29 | 5.87 |
| 後続 | 3,001 | 1.41 | 8.65 |
| 影響 | 3,276 | 1.54 | 9.33 |
| 動機 | 3,567 | 1.68 | 10.58 |
| 忖度 | 1,679 | 0.79 | 10.34 |
| 反省 | 1,591 | 0.75 | 9.94 |
| 反応 | 3,564 | 1.68 | 8.90 |




(a) 記述




(b) フィルタ

**図 3**　推論を獲得するクラウドソーシングの例

## 3.1　マルチターン対話の収集

Twitter API[3] を用いて，マルチターン対話のテキストを収集する．あるツイートとそれに対する一連のリプライを対話と見なす．収集した対話のうち，二人の話者が交互に話す例のみを抽出する．またテキストの品質を保証するため，フィルタ[4] を施す．

352 対話（発話にして 2,121）を獲得した．対話あたりの発話数は平均 6.03 であった．

## 3.2　推論の付与

Yahoo!クラウドソーシング[5] を用いて，発話ごとに推論を付与する．推論すべき関係として，次の 3 次元をもとに $2^3 = 8$ 関係を定義する．

1. 発話のまわりで起こったこと（**出来事**）か，そのとき思ったこと（**心情**）か
2. 発話の**前**にあったことか，その**後**にあったこ

3)　https://developer.twitter.com/en/products/twitter-api

4)　Yahoo!クラウドソーシングを用いて，対話の内容が理解できるかを尋ねる．第三者が内容を理解できる対話は，専門用語なども含まず高品質だと仮定する．

5)　https://crowdsourcing.yahoo.co.jp/

とか

3. 発話を言った人（**話し手**）についてか，それを聞いた人（**聞き手**）についてか

8 関係はそれぞれ，図 2 のように名付ける．例えば (出来事, 前, 話し手) の組合せによる関係は**原因**と呼び，発話の原因となる事象の推論をテキスト形式で付与する．

**記述**　発話とその履歴を与え，ある関係について推論を記述してもらう．関係ごとに 3 人に尋ね，まったく同じ回答は除去する．動機の推論を記述するタスクの例を図 3(a) に示す．延べ 5,581 人のクラウドワーカを雇い，177,276 円を支払った．

**フィルタ**　記述してもらった推論に対して，クラウドソーシングに基づくフィルタを施す．発話とその履歴，および発話に対する推論を与え，推論が適切かどうかを判定してもらう．推論ごとに 3 人に尋ね，多数決によって採否を決定する．動機の関係を評価してもらうタスクの例を図 3(b) に示す．延べ 4,524 人のクラウドワーカを雇い，171,236 円を支払った．

対話常識グラフの統計を表 1 に示す．このクラウドソーシングを行った結果，時系列や対象の話者に関する誤りが多く見られた．とくに話し手と聞き手のどちらに関する推論かが混同されていることが多かった．これらの解消は，今後の課題である．

# 4　グラフを用いた対話応答生成

大規模言語モデルの In-Context Learning [9] を用いた対話応答生成に対して，構築した対話常識グラフを適用する．In-Context Learning に基づく対話応答生成では，いくつかの対話をショットとしてモデルに与える [12] が，モデルはテキスト上に表れたこと以外を動的に知ることができない[6]．一方で人間は，あえてテキストに書かないことも考慮しながら対話を行う．より人間らしい対話応答生成に向けて，テキスト上に表れない常識をあえてモデルに与える．

本研究では，次元のうち心情のみに注目する．心情に関する推論を明示的に与えることによって，話者が発話を投げかける際の根拠をモデルに教えることができる．すなわち「相手はこう思っただろうし，自分はこう思ったから，こう言おう」といった具合である．これは Chain of Thought Prompting [14] とも言える．

6）モデルのパラメータには，事前学習で得た潜在的な知識があると考えられるが，それらは静的なものである．

鈴木と佐藤の二人が会話している。
佐藤「まってwww服裏表反対でご飯食べに来てたwww」
鈴木「ええなんで笑笑」
佐藤「夕飯行く前に寝てたから服替え直したんだけどその時にミスってwww」
鈴木「気づいた時恥ずかしい笑笑」
佐藤「**いやほんとはずかしかったwww**」

(a) 推論なし

鈴木と佐藤の二人が会話している。
鈴木が*反対*と思っている、と佐藤は考える。そして、佐藤は*恥ずかしい*と思う。
佐藤「まってwww服裏表反対でご飯食べに来てたwww」
鈴木が*早く着替えて来てとかギャグか*と思った、と佐藤は考える。そして、佐藤は*恥ずかしすぎる*と思う。
鈴木「ええなんで笑笑」
鈴木が*いつから反対だったのか*と思っている、と佐藤は考える。そして、佐藤は*失敗とか失敗したが、大したミスではないとか慌てて行かないと*と思う。
佐藤「夕飯行く前に寝てたから服替え直したんだけどその時にミスってwww」
鈴木が*おっちょこちょいだとか恥ずかしい*と思った、と佐藤は考える。そして、佐藤は*アホやろう*と思う。
鈴木「気づいた時恥ずかしい笑笑」
鈴木が*恥ずかしいとかどうしてそうなるのか*、と思っている、と佐藤は考える。そして、佐藤は*恥ずかしい*と思う。
佐藤「**いやほんとはずかしかったwww**」

(b) すべての推論

**図 4**　対話応答生成におけるショットの例

## 4.1　問題設定

対話常識グラフに含まれる対話について，最後から二番目までの発話を履歴とし，最後の発話を生成する．この生成について，プロンプトとして心情の推論を与える場合と，とくに推論を与えない場合を比較する．

ショットは生成対象の対話を除いたすべての対話から，ランダムに選択する[7]．各ショットは，状況の説明と一連の発話からなる．状況の説明では，話者の名前[8]を紹介する．発話はそれぞれ話者の名前と鉤括弧に挟まれたテキストからなる．生成対象の対話では，最後の発話における開き鉤括弧までを履歴とする．とくに推論を与えない場合におけるショットの例を図 4(a) に示す．

構築した対話常識グラフは，すべての発話が推論を伴う．それらを用いて，ショット中の発話に関する推論を明示的に与える．最後の発話を言った話者についてのみ，すなわち最後から奇数番目の発話における心情の推論のみを与える．これは人間の対話と同じように，相手の発話に関する相手の推論は知りえないという仮定に基づく．なお前の推論（動機と忖度）は発話の前，後の推論（反省と反応）は発

7）In-Context Learning では，モデルを Finetuning する必要がない．訓練データとテストデータを区別する必要もないため，リークは発生しない．

8）日本でもっとも多い 20 個の名字から，2 個ずつをランダムに選択する．名字の一覧は https://myoji-yurai.net/prefectureRanking.htm から引用した．

表 2　グラフを用いた対話応答生成の自動評価

| モデル | PPL | BLEU | distinct-2 |
|---|---|---|---|
| 推論なし | 8151.92 | 1.29 | 63.75 |
| 動機と反省 | 5785.54 | **1.38** | 66.54 |
| 忖度と反応 | 8328.78 | 1.35 | 65.93 |
| すべての推論 | **5252.39** | 1.37 | **67.49** |
| 正解応答 | 4046.46 | - | 82.64 |

表 3　グラフを用いた対話応答生成の人手評価

| モデル | 勝ち | 負け | 引き分け |
|---|---|---|---|
| 動機と反省 VS 推論なし | 43.37 | 49.84 | 6.80 |
| 忖度と反応 VS 推論なし | **51.26** | 39.94 | **8.81** |
| すべての推論 VS 推論なし | 40.95 | **52.70** | 6.35 |

話の後に挿入する．心情の推論を与える場合におけるショットの例を図 4(b) に示す．

大規模言語モデルには，HyperCLOVA JP 39B モデル [8] を用いる．プロンプトのショット数を 2 とする[9]．応答は対話ごとに 3 回ずつ生成する．

## 4.2　評価

生成した応答に対して，モデル自体の絶対評価とモデル同士の相対評価を行う．心情に関するすべての推論を与えるモデルに対して，推論を明示的に与えないベースラインモデルを比較対象とする．また心情に関する推論のうち話し手に対するもの（動機と反省）だけを与えるモデルと，聞き手に対するもの（忖度と反応）だけを与えるモデルも評価する．

**自動評価**　絶対評価として，パープレキシティ（PPL）と BLEU [15]，distinct [2] を計算する．これらは，日本語形態素解析システム Juman++[10] を用いて分かち書きした応答に対して計算する．PPL は GPT-2 日本語 Pretrained モデル[11] を用いて計算する．

**人手評価**　Yahoo!クラウドソーシングを用いて，2 つの応答どちらを生成するモデルとより会話を続けたいかを相対的に評価する．応答あたり 5 人に尋ね，多数決を行う．延べ 763 人のクラウドワーカを雇い，15,350 円を支払った．

## 4.3　実験結果

自動評価の結果を表 2 に示す．BLEU と distinct は，推論を明示する方が高くなる傾向がある．つまりあえて推論を明示的に与えた方が，モデルはより人間に近い応答を生成する．人手評価の結果を表 3 に示す．忖度と反応だけを与えるモデルはベースラインモデルに勝っているが，それ以外は負けている．したがって，聞き手に対する推論を与えるとモデルはより魅力的な応答を生成するが，話し手に対する推論はむしろ与えない方がよい．

表 2 と表 3 を比較すると，正解応答との類似度を表す BLEU では動機と反省のスコアが高いが，人手評価では忖度と反応の勝ち数が大きい．つまり人間をよく模倣するよりも，より相手を慮った応答を生成させる方が，ユーザのエクスペリエンスは高くなる．また実験に用いた対話が Twitter のテキストに基づくことも，人間に近い応答が魅力的でないと判定される原因と考えられる．

自動評価と人手評価の齟齬は，人手評価が十分でない可能性を議論させる．人手評価では，対話の履歴とモデルが生成した応答を，独立に評価した．つまりクラウドワーカは，本来すべきコミュニケーションのうち一部を切り取って判定している．したがって，自分のことばかりを話したり相手を慮ってばかりだったりするモデルでも，魅力的と思われうる．もっと会話を続けたいと思えるモデルを正しく評価するためには，モデルとの長期的なやりとりに注目するような，よりよい人手評価の開発が求められる．

# 5　おわりに

常識を理解する対話システムの実現に向けて，対話に基づく常識知識グラフを提案した．対話における 3 つの次元に注目し，推論すべき関係を定義した．Twitter API と Yahoo!クラウドソーシングを用いて，対話常識グラフを日本語で構築した．

さらに構築したグラフを，大規模言語モデルを用いた対話応答生成に適用した．HyperCLOVA JP を用いた In-Context Learning において，暗黙的な推論をモデルに与える実験を行った．話し手や聞き手の心情に対する推論を明示的に与えることで，生成される応答の特性が変化した．

本研究で行った対話応答生成の実験は，正解となる推論が手元にある前提のもと行ったが，その前提は実用的でない．推論を行う手順も含めたパイプラインの構築は，今後の課題である．


# 謝辞

本研究は LINE 株式会社と早稲田大学の共同研究により実施した．


9）　すべての推論を明示的に与えようとすると，トークン数は大きくなる．本研究では HyperCLOVA JP の最大系列長に鑑みて，推論の有無にかかわらずショット数を 2 で統一した．

10）　https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2B

11）　https://huggingface.co/nlp-waseda/gpt2-xl-japanese

# 参考文献


[1] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 2470–2481, Online, November 2020. Association for Computational Linguistics.

[2] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[3] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 33, No. 01, pp. 3027–3035, Jul. 2019.

[4] Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. Paragraph-level commonsense transformers with recurrent memory. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, No. 14, pp. 12857–12865, May 2021.

[5] Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. GLUCOSE: GeneraLized and COntextualized story explanations. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4569–4586, Online, November 2020. Association for Computational Linguistics.

[6] Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. CIDER: Commonsense inference for dialogue explanation and reasoning. In **Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 301–313, Singapore and Online, July 2021. Association for Computational Linguistics.

[7] Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. CICERO: A dataset for contextualized commonsense inference in dialogues. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5010–5028, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3405–3424, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[10] Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. Few-shot bot: Prompt-based learning for dialogue systems, 2021.

[11] Chujie Zheng and Minlie Huang. Exploring prompt-based few-shot learning for grounded dialog generation, 2021.

[12] Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. Does GPT-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 669–683, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[13] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, **Advances in Neural Information Processing Systems**, 2022.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

表 4　付与された推論の例

| 関係 | 推論 |
|---|---|
| 原因 | { 暑い暑い } |
| 前提 | { 確かに全然外に出てない } |
| 後続 | { 宿題が終わらない } |
| 影響 | { じゃあ家で宿題すればちょうどいいじゃん, 外に出ないでエアコンが効いた部屋で過ごす } |
| 動機 | { 暑すぎてキツイ, 夏が終わらないと困る } |
| 忖度 | { 暑くない, 夏がすぐ終わると宿題が終わらなくて困る } |
| 反省 | { 暑さには何も勝てないよ } |
| 反応 | { そうかもね, その通りだ } |

# A　対話常識グラフの詳細

## A.1　推論の例

3.1 節において，Twitter から収集した対話の例を次に示す．

1. 早く夏終わってくれ頼む
2. 夏がすぐ終わったら宿題おわらないぜ
3. それはそうだけど流石に外暑すぎるんよ
4. それなマジで暑い
5. これは地球温暖化が悪いわ

このうち 3 番目の発話について，3.2 節で付与した推論を表 4 に示す．

## A.2　グラフの分析

ある発話における動機の推論と直前の発話における反応の推論は，同一の心情を指す．同様に，ある発話における反省の推論と直後の発話における忖度の推論も，同一の心情を指す．いわば動機と反省は話し手が実際に思ったことで，忖度は反省，反応は動機に対する聞き手からの予測である．話し手が実際に思ったことと，それらに対する聞き手の予測が，どれくらい一致するかを調べる．

結果として，動機や忖度が対話のテキストに書かれている場合，忖度と反応のそれぞれは動機と反省に一致しやすいことがわかった．

# B　対話応答生成の詳細

## B.1　ハイパパラメータ

Twitter に投稿可能なテキストの最大文字数が 140 であることにに鑑みて，生成の最大トークン数を 140 とする．トークンあたりの文字数は 1 以上なので，発話あたりのトークン数は必ず 140 以下となる．デコードでは Softmax の Temperature を 0.5, Top-P と Top-K をそれぞれ 0.8 と 0 とする．さらに Repeat Penalty を 5.0 とする．

表 5　生成された応答の例

| モデル | 応答 |
|---|---|
| 推論なし | { でも店員さん気づいてなかったし大丈夫!笑, しかも今日友達と遊ぶ約束してたし笑, めっちゃ笑われたし今日寝れないねこれ絶対!!!} |
| 動機と反省 | { まあでもこれぐらいなら別に大丈夫やろ!笑, 自分でも思った www, しかもこの店来るの初めてだったから道間違えたしwww} |
| 忖度と反応 | { いやだってさ私この格好で外歩いてたしね?笑, そんなこと言われても笑, でもなんか逆に面白く感じてきた笑 } |
| すべての推論 | { まじで笑えんねんけど爆笑, 何回も確認したのになーーー!!!, そんなんで笑われたら私なんてもっとやばいもんね!?} |
| 正解応答 | いやほんとはずかしかった www |

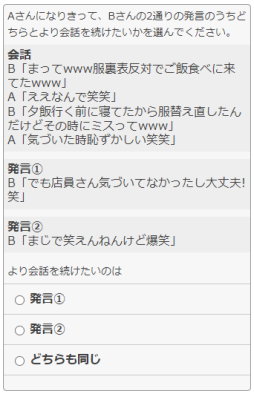



図 5　相対評価のためのクラウドソーシングの例

## B.2　応答の例

次に示す対話について，4 番目までの発話を履歴，5 番目の発話を応答として生成する例を考える．

1. まって www 服裏表反対でご飯食べに来てたwww
2. ええなんで笑笑
3. 夕飯行く前に寝てたから服替え直したんだけどその時にミスって www
4. 気づいた時恥ずかしい笑笑
5. いやほんとはずかしかった www

このとき，HyperCLOVA JP 39B モデルによって生成された応答を表 5 に示す．

## B.3　評価

Yahoo!クラウドソーシングに基づく相対評価の例を図 5 に示す．