2022 年度 修士論文

感情と常識を理解する対話システム開発のための日本語データセット構築

Building Japanese Datasets for Emotion and Commonsense-Aware Dialogue Systems

指導教員　河原 大輔 教授

研究指導名　自然言語処理研究

早稲田大学 基幹理工学研究科 情報理工・情報通信専攻

学籍番号 5121F011

井手 竜也

2023 年 1 月 23 日

概要

コンピュータとのコミュニケーションは，人間の生活を豊かにする．社会の情報化もあいまって，その需要は高まっている．深層学習の発展をもとに，自然言語処理の発展が著しい．人間と雑談する対話システムも盛んに開発されている．対話において，人間は感情を考慮したり常識を推論したりする．一方で，コンピュータが暗黙的な実世界知識を理解するのは難しい．人間のように，コンピュータが感情や常識を理解するための研究がなされている．多くのデータセットが構築されているが，日本語のものは少ない．本研究では，感情や常識を理解する対話システムの開発に向けた日本語データセットを提案する．話者が抱く感情やテキストに書かれない常識に注目し，構築すべきデータセットをデザインする．対話に基づく感情と常識をキーワードに，発話の話し手と聞き手が抱く感情を分けてタグ付けした対話コーパス・イベントとメンタルステートに関する推論を収集した常識知識グラフ・対話中の発話文における暗黙的な推論を収集した常識知識グラフを構築する．クラウドソーシングと大規模言語モデルを用いて，それらを構築する手法を提案する．データセットを分析した結果，対話における感情や常識の傾向を明らかにした．またデータセットを用いた実験から，人間の特性や人間とコンピュータの違いを明らかにした．構築したデータセットは公開予定である．

目 次

図 目 次

表 目 次

1 はじめに

人間がコンピュータと関係を築くことは、人間の生活をより豊かにする。社会の情報化やオンライン化が進み、人間とコンピュータのコミュニケーションに対する需要はますます高まっている。計算資源の拡大によって、深層学習に基づく自然言語処理が著しく発展している。ニューラル言語モデルを用いた対話システムも、盛んに開発されている。タスク志向型のものから、オープンドメインな雑談ができるものまで存在する。

人間は対話において、相手の感情を考慮したり自分の感情を表現したりする。また物事を推論したり質問に答えたりするとき、人間は常識を用いる。一方、コンピュータに感情や常識といった実世界知識を与えることは難しい。とくに深層学習に基づくニューラル言語モデルは、テキストに表れない暗黙的な常識を捉えられない。コンピュータに感情や常識を理解させるため，多くのデータセットが構築されている。例えば、発話ごとに感情や対話行為といった特徴をタグ付けした対話コーパス [Li 17, Rashkin 19] が存在する。他にも、概念や常識を関係で繋いだ知識グラフ [Speer 17, Sap 19, Hwang 21] が存在する。しかしそれらは基本的に英語で、日本語のものはほとんどない。

本研究では、対話システムの常識理解を促進するデータセットを日本語で構築する。対話における感情と常識をキーワードに、以下のデータセットを提案する。

1. ある発話において、話し手と聞き手が抱く感情 [Buechel 17] を分けてタグ付けした感情タグ付き対話コーパス

2. あるイベントの前後において、その人に起こるイベントあるいはその人が思うメンタルステート [Rashkin 18, Sap 19] を集めた常識知識グラフ

3. 対話中の発話文それぞれに対して，イベントとメンタルステートに関する暗黙的な推論 [Ghosal 22] を書いた常識知識グラフ

データセットを日本語で構築するだけでなく、タグ付けにおける観点を拡張したり構築にかかるコストを下げたりと、既存のデータセットがもつ問題点の解決も図る。

Yahoo!クラウドソーシング[1]と大規模言語モデルの HyperCLOVA JP [Kim 21] を用いて、提案するデータセットを日本語で構築した。感情タグ付きマルチターン対話コーパスでは、Twitter から収集したマルチターン対話コーパスに対して，Yahoo!クラウドソーシングによって発話ごとに話し手と聞き手が抱く感情をタグ付けした。イベント常識知識グラフでは、Yahoo!クラウドソーシングと HyperCLOVA JP を併せて、ゼロからの低コストな構築を試みた。Yahoo!クラウドソーシングによって構築した小規模なグラフを、HyperCLOVA JP を用いた In-Context Learning [Brown 20] によって大規模に拡大 [West 22] した。対話常識グラフでは、感情タグ付き対話コーパスと同じように Twitter から収集したマルチターン対話コーパスに対して、Yahoo!クラウドソーシングによって発話ごとに推論を付与した。付与に際して、カテゴリや時系列、対象人物といった次元を考慮し、推論すべき関係を定義した。

構築した感情タグ付き対話コーパスの分析から、ある発話において話し手と聞き手が抱く感情は異なることや、それらの間には相関があることがわかった。また BERT [Devlin 19] を用いた感情認識の実験では、話し手よりも聞き手の感情を予測する方が難しいことを明らかにした。イベント常識知識グラフの構築では、Yahoo!クラウドソーシングのタスクと HyperCLOVA

[1] https://crowdsourcing.yahoo.co.jp/

JP による生成、すなわち人間と言語モデルに対するプロンプティングで推論の傾向が異なることを示した。さらに対話に基づく常識知識グラフに関して、HyperCLOVA JP の In-Context Learning を用いた対話応答生成に対してそれを適用した。暗黙的な推論を明示的に入力することで、生成される応答の特性が変化することを示した。

なお本研究で提案した日本語のデータセットを構築する手法は、ほかの言語に対して適用することも可能である。本研究で構築したデータセットが、日本語を始めとするコンピュータの常識理解を促進することを期待する。構築したデータセットはすべて、公開予定である。

2 関連研究

2.1 感情タグ付きコーパス

EmoBank [Buechel 17] はある文を書いた人とそれを読んだ人の感情を、クラウドソーシングによってタグ付けしたコーパスである。WRIME [Kajiwara 21] は EmoBank と同様、主観的と客観的な感情をタグ付けした日本語のコーパスである。ただし EmoBank よりも主観性と客観性に重きを置いており、投稿者自身に主観的な感情をタグ付けさせている。EmoInt [Mohammad 17] は怒り・恐れ・喜び・悲しみの 4 感情について、その強度をタグ付けしたコーパスである。

文の感情だけでなく、その原因にも注目するようなコーパスもある。Emotion Stimulus [Ghazi 15] は FrameNet のフレームを用い、文の感情とその原因をタグ付けしたコーパスである。ラベルとしては Ekman の 6 感情 [Paul 92] に恥ずかしさを加えた 7 感情を採用している。一方で、Grounded Emotions [Liu 17] は Twitter のツイートと天気やニュースの関係をアノテーションしている。こちらはラベルに喜びと悲しみのみを採用している。

StoryCommonsense [Rashkin 18] は短い物語をなす一連の文に、登場人物の Motivation と Emotional Reaction をタグ付けしたコーパスである。この感情タグは Plutchik の 8 感情 [Plutchik 80] に基づく。しかし、上記のコーパスはどれも対話に関するものではない。

2.2 対話コーパス

EmpatheticDialogues [Rashkin 19] は 2 人の話者を Speaker と Listener に割り当て、対話ごとに Speaker の感情とその状況をタグ付けした対話コーパスであり、日本語版 [Sugiyama 21] も存在する。EmpatheticDialogues は対話ごとに感情をタグ付けしているため、発話ごとの感情を理解することには向かない。DailyDialog [Li 17] は発話ごとに感情と意図をタグ付けした対話コーパスである。EmotionLines [Hsu 18] は発話ごとに感情をタグ付けした対話コーパスで、話者が複数人である。これらはともに Ekman の 6 感情に無感情もしくはその他を加えた 7 個のラベルを用いている。本研究では DailyDialog や EmotionLines と同じく、発話ごとに感情をタグ付けする。これらの対話コーパスは発話に込められた感情のみをタグ付けしているが、本研究ではその発話を聞いた相手が抱く感情もタグ付けする。さらに本研究では複数の感情を許容し、それぞれの感情に強弱を設ける。

2.3　常識知識グラフ

ConceptNet [Speer 17] は単語や句を関係でつないだ知識グラフで、とくにエンティティ同士の関係に注目している。ATOMIC [Sap 19] はイベント同士の関係に加え、イベントとメンタルステートの関係を扱う知識グラフである。これらの知識はクラウドソーシングを通して、人の手によって書かれている。ATOMIC と ConseptNet をマージして拡張した知識グラフに、ATOMIC-2020 [Hwang 21] がある。

ASER [Zhang 19] は文の談話関係に注目し、テキストデータから抽出された知識グラフである。アクティビティとステート、イベント間の関係を扱う。さらに TransOMCS [Zhang 20a] は、ASER のような知識グラフからパターンマッチやランキングによってブートストラップ的に新たな常識を発見するアプローチである。

ConceptNet や ATOMIC はクラウドソーシングによって獲得されているのに対して、ASER と TransOMCS はともに自動獲得である。自動獲得の場合、大規模な構築は容易だが、テキストに現れない知識を得ることは難しい。一方、クラウドソーシングであれば良質なデータを集められるが、金銭的にも時間的にもコストが高い。

QA のような形式で常識を収集したデータセットもある。COPA [Roemmele 11] では premise に対して plausible な cause と effect を、二択の質問として獲得している。SWAG [Zellers 18] ではビデオのキャプションから、ある状況の記述に対する推論を四択の質問として獲得している。KUCI [Omura 20] は SWAG のような日本語の常識推論データセットで、automatic extraction と crowdsourcing の併用から構築されている。CommonsenseQA [Talmor 19] はコンセプトに関する常識を QA として扱ったデータセットで、ConceptNet から構築されている。

知識をシンボリックではなくニューラルに蓄える研究があり、ニューラル言語モデルを知識ベースとして扱う手法 [Petroni 19, AlKhamissi 22] が注目されている。COMET [Bosselut 19] は事前学習済み Transformer を ATOMIC と ConceptNet で Fine-Tuning し、見たことがないイベントに対する推論を目指す。PARA-COMET [Gabriel 21] は Memory Component を導入し、段落レベルの情報を考慮した推論を目指す。

大規模言語モデルがもつ常識をより小さな言語モデルに蒸留する試み [West 22] がある。そこでは GPT-3 [Brown 20] を用いて ATOMIC を拡大し、RoBERTa [Liu 20] を用いてフィルタを施している。

特定の文脈に基づく常識知識グラフもある。GLUCOSE [Mostafazadeh 20] は物語の文に関する推論、CIDER [Ghosal 21] は対話における発話同士の推論を扱う。CICERO [Ghosal 22] 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。発話ごとに複数の推論をタグ付けした、CICEROv2 [Shen 22] もある。

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。

3　表出感情と経験感情をタグ付けした対話コーパスの構築

相手の感情を認識した上で適切な感情を込めた応答を生成できる対話システムを実現するための感情タグ付き対話コーパス[2]を構築する。話し手が発話に込めた感情（**表出感情**）と発

[2] https://github.com/nlp-waseda/expr-exper-emo にて公開している。

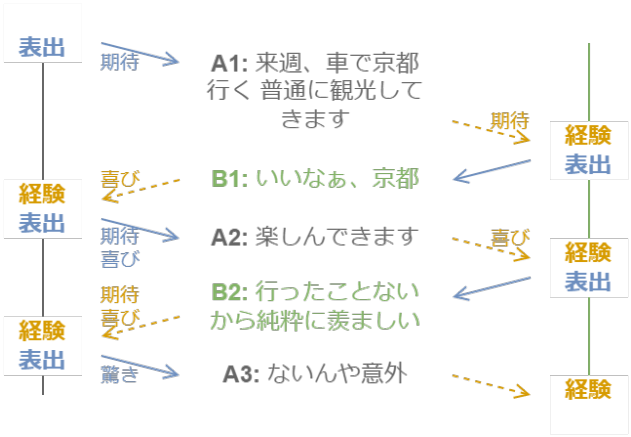


図 1: 表出感情と経験感情のタグ付け

発話	表出		経験	
	強	弱	強	弱
A1: 来週、車で京都行く 普通に観光してきます	{期待}	{期待, 喜び}	{期待}	{期待}
B1: いいなぁ、京都	{}	{期待}	{喜び}	{期待, 喜び}
A2: 楽しんできます	{期待, 喜び}	{期待, 喜び}	{喜び}	{期待, 喜び}
B2: 行ったことないから純粋に羨ましい	{}	{期待}	{期待, 喜び}	{期待, 喜び}
A3: ないんや意外	{驚き}	{驚き}	{}	{喜び, 驚き}

表 1: 表出感情と経験感情がタグ付けされた対話の例

話の聞き手が受けた感情（**経験感情**）を発話ごとにタグ付けする。なお本研究ではこれを日本語で構築するが、手法はどの言語にも適用できる。

3.1　コーパスの構築

3.1.1　対話の収集

対話は Twitter API によって収集する。2 人のユーザによるツイートとリプライのかけあいを対話とみなし、それを抽出する。ただしハッシュタグや URL、画像を含む対話はすべて除外する。対話あたりの発話数は 2 から 9 までに絞り、いくつかのフィルタを施す。まず特殊な記号や絵文字を含む対話はすべて除外する。ほかにも 4 回以上繰り返される文字や単語を含む場合や、発話が 4 文字未満の場合も除外する。

上記の手法によって、3,828 対話（13,806 発話）を獲得した。ただしターン数が多い対話ほど、そのサンプル数は減る傾向にある。

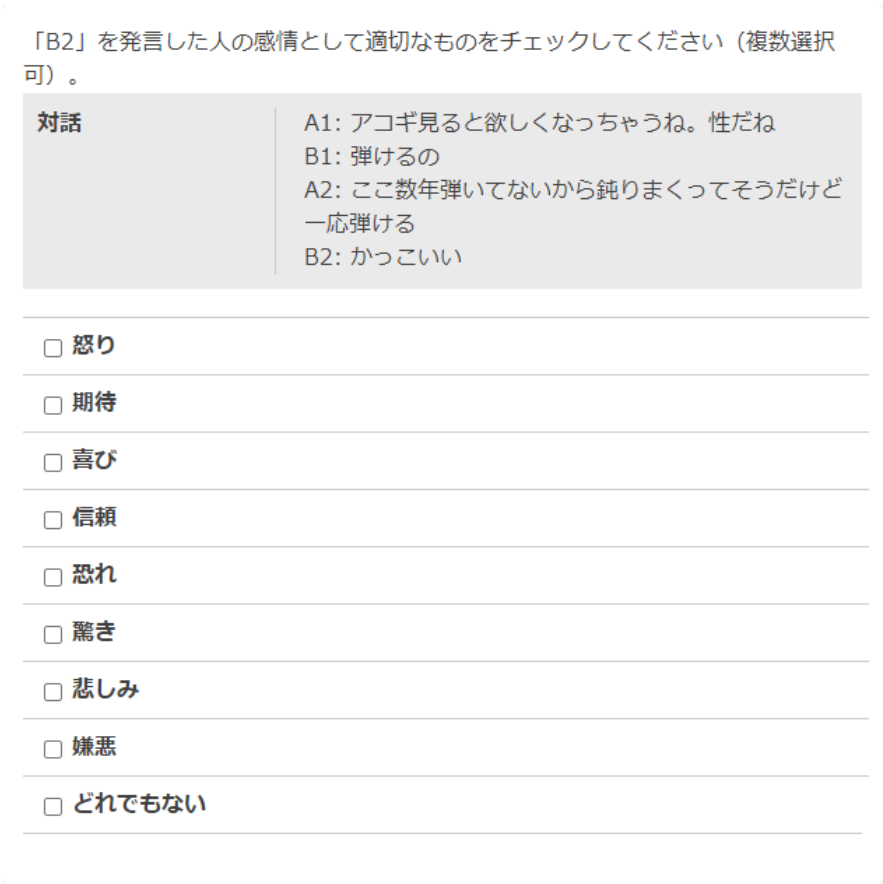


図 2: クラウドソーシングの例

3.1.2　表出感情と経験感情のタグ付け

本研究では Plutchik の感情の輪 [Plutchik 80] をラベルに採用する。[3]具体的には{怒り, 期待, 喜び, 信頼, 恐れ, 驚き, 悲しみ, 嫌悪}の 8 感情である。発話ごとにそれを話した人が抱いていた感情（表出感情）とそれを聞いた人が抱いた感情（経験感情）をタグ付けする。つまり発話ごとに主観的な感情と客観的な感情 [Buechel 17, Kajiwara 21] をタグ付けすることになる。表出感情と経験感情を分けてタグ付けすることで、発話間や話者間における感情の変化を捉える。

感情のタグ付けはクラウドソーシングによって行う。プラットフォームとしては Yahoo!クラウドソーシングを用い、クラウドワーカに発話とその履歴を与え、各感情の有無を問う。複数の感情を選択することや、どの感情も選択しないことを許容する。発話ごとに 7 人のクラウドワーカを雇い、半数（4 人）以上が選んだ感情を強ラベル、4 分の 1（2 人）以上が選んだ感情を弱ラベルとする。ここで強ラベルの集合は弱ラベルの集合の部分集合となる。このタグ付けを表出感情と経験感情のそれぞれについて独立に行う。表出感情をタグ付けするためのクラウドソーシング画面の例を図 2 に示す。チェックボックス式のフォーマットを用いることによって、複数の感情が選ばれることを許容している。タグ付けされた対話の例を表 1 に示す。

[3]感情タグ付きコーパスでは、Ekman の 6 感情 [Paul 92] や Plutchik の感情の輪 [Plutchik 80] がよく用いられる。クラウドソーシングを用いた予備実験の結果、後者の方がより感情タグとして適切であることが明らかになった。したがって本研究ではそれを採用する。

ラベル	表出		経験	
	強	弱	強	弱
怒り	430	1,349	124	870
期待	1,906	4,229	1,215	4,068
喜び	1,629	3,672	1,553	4,549
信頼	247	1,732	520	3,455
恐れ	252	942	123	846
驚き	602	2,018	434	2,798
悲しみ	1,227	2,936	889	3,037
嫌悪	476	1,979	186	1,535
どれか	6,371	12,215	4,705	12,515

表 2: ラベルごとの発話数

ラベルごとの発話数を表 2 に示す。表出感情に関して、少なくとも 1 個の強ラベルを伴う発話は 46.15%、弱ラベルのそれは 88.48%である。経験感情ではそれぞれ 34.08%と 90.65%である。およそ 9 割の発話がいずれかの弱ラベルを伴っており、その他の感情や無感情のラベルが大部分を占める DailyDialog [Li 17] や EmotionLines [Hsu 18] よりも有効である。

3.2　コーパスの分析

3.2.1　感情ごとに頻出な単語

感情ごとの発話がもつ特性を知るため、感情ごとに単語の頻度を求める。本研究では強ラベルのみを対象とし、品詞として動詞と形容詞を採用する。品詞の解析は日本語形態素解析システム Juman++ [Tolmachev 18] によって行い、代表表記を抽出する。どの感情にも現れる普遍的な単語を除くため、IDF によるフィルタを施す。具体的には IDF が最大値の半分を下回った単語を無視する。

フィルタを施した、感情ごとに頻出な単語の Top-3 を表 3 に示す。喜びの動詞や喜び・恐れ・嫌悪の形容詞は表出感情と経験感情で等しい。一方で、怒り・嫌悪の動詞や信頼の形容詞は表出感情と経験感情で異なる。喜びや驚きの動詞に「交じる」があるが、これは形容詞の「マジだ」が誤って識別されたものである。同様に、喜びの形容詞にある「孵る」は「帰る」に対する誤りである。

3.2.2　表出感情と経験感情の関係

構築した対話コーパスにおける、表出感情と経験感情の関係を分析した。具体的には、次のような関係が考えられる。

1. 同一の発話に対する表出感情と経験感情（異なる話者）

2. ある発話に対する経験感情と次の発話に対する表出感情（同一の話者）

表 3: ラベルごとに頻出な単語の Top-3

ラベル	表出		経験	
	動詞	形容詞	動詞	形容詞
怒り	止める, 許す, 違う	マジだ, うるさい, 馬鹿だ	話す, 広げる, 居る	うるさい, マジだ, 悪い
期待	教える, 願う, 待つ	一緒だ, 楽しみだ, 面白い	待つ, 教える, 上げる	楽しみだ, 強い, 一緒だ
喜び	笑う, 交じる, 孵る	楽しい, 嬉しい, おもろい	笑う, 交じる, 孵る	楽しい, 嬉しい, おもろい
信頼	飲む, 願う, らっしゃる	大丈夫だ, 優しい, 一緒だ	教える, 付ける, 待つ	大事だ, 大丈夫だ, 同じだ
恐れ	付ける, 知れる, 助ける	怖い, やばい, 危険だ	知れる, 付ける, 入る	怖い, やばい, 危険だ
驚き	交じる, 気付く, 居る	やばい, 怖い, 強い	居る, 気付く, 知る	怖い, 痛い, やばい
悲しみ	泣く, 生きる, 帰る	痛い, 悲しい, 辛い	泣く, 生きる, 消える	辛い, 痛い, 悲しい
嫌悪	知る, 止める, 交じる	悪い, 嫌いだ, 嫌だ	困る, 上げる, 働く	悪い, 嫌だ, 嫌いだ

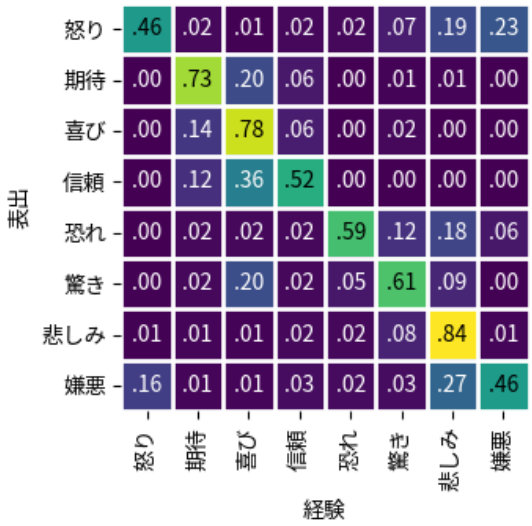
強ラベルに関する、それぞれの混同行列を図 3 に示す。なお各要素の値は行方向に正規化してある。対角成分における値が大きくなっていることから、同一の発話および同一の話者についての関係としては、同じ感情が起こりやすいことが言える。図 3(a) から、期待・信頼・驚きの発話に対して喜びを抱きやすいことがわかる。また怒りと嫌悪にはそれぞれ嫌悪と悲しみを抱きやすい。一方で図 3(b) を見ると、信頼を抱いたあとは信頼よりも喜びを表しやすいことがわかる。怒りに関しても、そのあとは怒りよりも嫌悪を表しやすい。また図 3(b) から、信頼と怒りを抱いたあとはそれぞれ喜びと嫌悪を表しやすいことがわかる。図 3(a) と図 3(b) を比較したとき、とくに悲しみにまつわる関係が異なっている。特定の発話では、悲しみの発話が相手に悲しみを抱かせる。しかし同一話者間では、話者が悲しみを抱いたときに期待を表すことがある。これは相手の発話に対して悲しみを抱いた話者が、相手を慰めるために期待感のある発話を返していると考えられる。

3.3　感情認識の実験

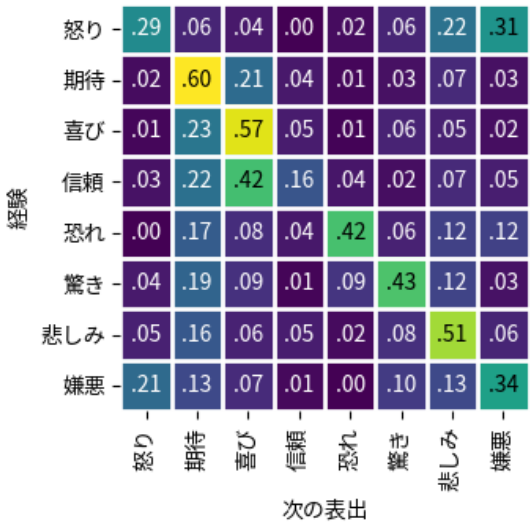
3.3.1　モデル

表出感情と経験感情について、それらの認識を実験する。対象の発話とそれまでの履歴から、各ラベルの強度を回帰することを考える。ラベルがない場合に 0、弱ラベルと強ラベルにそれぞれ 1 と 2 を充てる。モデルは平均二乗誤差で学習する。

本研究ではモデルとして BERT [Devlin 19] を利用する。事前学習済み BERT として、京都



(a) 表出感情と経験感情（特定の発話について）



(b) 経験感情と次の表出感情（特定の話者について）

図 3: 表出感情と経験感情の関係

モデル	表出		経験	
	Pearson	Spearman	Pearson	Spearman
京大 (BASE)	58.84	44.33	53.60	41.84
京大 (LARGE)	60.85	45.16	55.09	42.94
NICT	**61.50**	**46.05**	**56.23**	**43.88**

表 4: モデルごとの相関係数

大学 [柴田 19] の WWM 版と NICT[4]の BPE 版を採用する。入力は日本語形態素解析システム Juman++ [Tolmachev 18] によって単語に分割し、さらに BPE [Sennrich 16] によってサブワードに分割する。履歴を含む各発話を [SEP] で結合し、先頭と末尾にそれぞれ [CLS] と [SEP] を付加する。ここで対話の話者は 2 人であるから、それぞれによる発話に BERT の Segment ID として 0 または 1 を与える。これによってモデルに話者の情報を与える。[CLS] に対応するベクトルを全結合層に与え、Plutchik の 8 感情に対応する 8 次元のベクトルを得る。このベクトルがもつ各要素をそれぞれ回帰することになる。

対話コーパスを 8:1:1 に分割し、それぞれ学習データ・検証データ・テストデータに充てる。学習は 3 エポックと定め、テストデータに対して評価を行う。評価指標には Pearson と Spearman の相関係数を採用する。

3.3.2　結果

モデルごとの相関係数を表 4 に示す。モデルとしては、どの値も NICT がもっとも高い。表出感情と経験感情を比較すると、どのモデルも経験感情の値が低い。つまりある発話を話した人が抱いていた感情よりも、それを聞いた人が抱いた感情を予測することの方が難しい。

[4] https://alaginrc.nict.go.jp/nict-bert/index.html

ラベル	表出		経験	
	Pearson	Spearman	Pearson	Spearman
怒り	50.21	33.80	38.11	23.80
期待	62.76	**55.55**	57.46	51.22
喜び	**67.25**	55.22	**61.92**	**54.47**
信頼	41.15	36.69	43.91	40.48
恐れ	59.09	31.47	49.60	24.90
驚き	49.86	39.58	40.58	33.86
悲しみ	63.70	51.50	55.48	43.88
嫌悪	47.76	38.18	37.32	28.13

表 5: ラベルごとの相関係数 (NICT)

対話	ラベル	
	予測	正解
A1: ゲームの検証してる人が検証してほしいことあれば言ってください的なこと言ってたから依頼したら無視されて悲しくなったのはいい思い出 B1: それは悲しいね	**強い悲しみ**	**強い悲しみ**
A1: youtube でバーのマスターが氷砕いてる動画見てボーッとしてる B1: なんかしてよ, そのうちこういうときにツイキャスをしようかなと思っておる A2: 天才の発想 スマホでも見やすいから助かる	弱い期待 弱い**喜び**	強い**喜び** 弱い信頼
A1: 今、部活終わって帰るとこやけど 雨やばいしかっぱ持ってきてないし 最悪 B1: わたしも学校出た瞬間大雨降ってきた	強い驚き	強い悲しみ

表 6: 予測結果の例（NICT、最後の発話に対する表出感情）

NICT のモデルに関する、ラベルごとの相関係数を表 5 に示す。表 5 と表 2 を比較すると、ラベルの発話数が多いほど相関係数の値も高くなっていることがわかる。モデルによる予測結果の例を表 6 に示す。

3.3.3　マルチタスク学習

単一のモデルに表出感情と経験感情の両方を認識させることを考える。表出感情と経験感情には相関があるため、このことはそれぞれの認識に良い影響があると考えられる。タスクごとに全結合層を用意し、それらを同時に訓練する。ロスはそれぞれに対して計算し、その算術平均を全体のロスとしてパラメータを最適化する。

$$L_{\text{multi-task}} = \frac{L_{\text{expressed}} + L_{\text{experienced}}}{2} \tag{1}$$

3.2 章の関係を踏まえ、特定の発話と特定の話者に関する表出感情と経験感情をマルチタスク学習 [Liu 19] する。また訓練データとテストデータが異なる場合も実験する。マルチタス

訓練\テスト	表出		経験	
	Pearson	Spearman	Pearson	Spearman
表出	61.50	46.05	52.89	40.91
経験	55.49	43.34	56.23	43.88
マルチタスク	**62.20**	**46.63**	**57.35**	**45.01**

表 7: 特定の発話に関するマルチタスク学習

訓練\テスト	経験		次の表出	
	Pearson	Spearman	Pearson	Spearman
経験	54.62	43.47	29.53	25.46
次の表出	43.32	35.27	33.91	28.31
マルチタスク	**55.75**	**49.50**	**35.17**	**30.49**

表 8: 特定の話者に関するマルチタスク学習

ク学習の相関係数を表 7 と表 8 に示す。まず訓練とテストでデータが異なる場合、それらが同じ場合よりも値が低い。これは表出感情と経験感情を分けてタグ付けすることに意味があることを示している。マルチタスク学習としては、表出感情と経験感情のどちらもそれらを同時に学習させたときの方が値は高い。つまり表出・経験・次の表出感情は互いに助け合うと言える。

4 大規模言語モデルを用いたイベント常識知識グラフの構築

クラウドソーシングと大規模言語モデルによってゼロから知識グラフを構築する手法を提案する。知識をクラウドワーカに記述してもらうのと大規模言語モデルに生成させるのは、どちらも少しの例から多くの例を作成してもらうという意味で、本質的に同じことである。異なるのは人間か言語モデルかだけで、いわば後者は前者のアナロジーである。本研究では、人間と言語モデルに対するプロンプトを用いて、段階的に知識グラフを構築する手法を提案する。指示と回答例をプロンプトと見なすクラウドソーシングによって小規模な知識グラフを獲得し、その一部をショットとして大規模言語モデルに大規模な知識グラフを生成させる。

4.1 知識グラフの構築

本論文では、クラウドソーシングと大規模言語モデルを用いて、常識推論に関する知識グラフをゼロから構築する手法を提案する。まずクラウドソーシングによって小規模な知識グラフを構築し、それをプロンプトに用いることによって、大規模言語モデルがもつ知識を抽出する。提案手法の概要を図 4 に示す。クラウドソーシングだけを用いてゼロから知識グラフを構築する場合、金銭的にも時間的にもコストが高い。クラウドソーシングと超大規模言語モデルを併用することで、とくに時間的なコストの削減が期待される。

ATOMIC [Sap 19] や ASER [Zhang 19] のような、イベント関する知識グラフを日本語でゼロから構築する。一段階目のクラウドソーシングには Yahoo!クラウドソーシング、二段階目

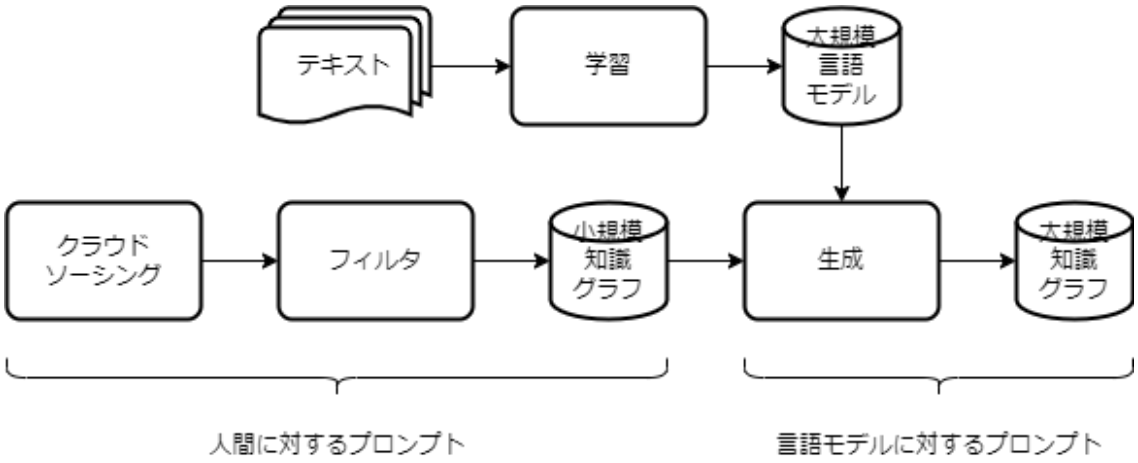


図 4: 提案手法の概要

	個数	適切	適切 [%]	Fleiss's κ
イベント	257	-	-	-
必要	504	402	79.76	39.85
影響	621	554	89.21	25.00
意図	603	519	86.07	36.11
反応	639	550	86.07	31.82

表 9: 小規模な知識グラフの統計

の超大規模言語モデルには HyperCLOVA JP [Kim 21] を用いる。

4.1.1　クラウドソーシングによる収集

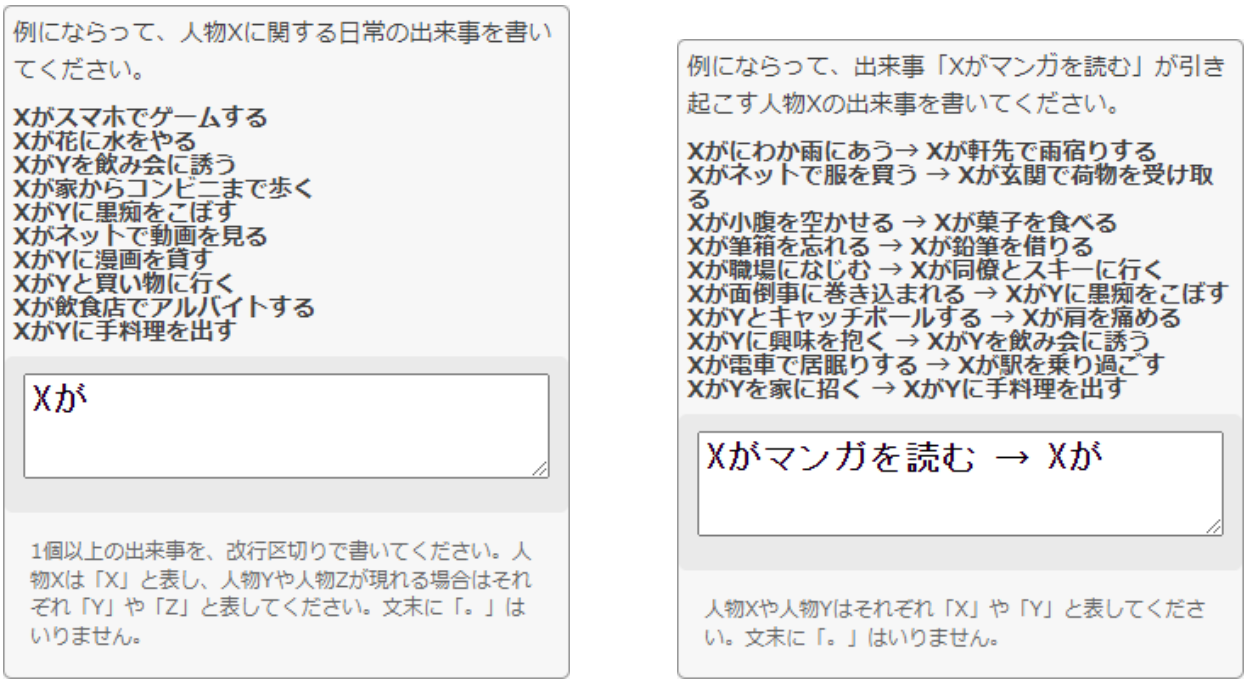
まずはイベントだけを集め、続いてそれぞれのイベントに対する推論を集める。

イベント　ある人物 X や周辺の人物 Y、人物 Z に関する日常的なイベントをクラウドワーカに記述してもらう。タスクの例を図 5(a) に示す。タスクでは指示と 10 個の例を与え、1 人あたり 1 個以上のイベントを記述してもらう。重複するイベントは集計時に取り除く。

推論　獲得したイベントに対して、その前後に対する推論をクラウドワーカに記述してもらう。推論する関係には、ATOMIC のそれにならって以下の 4 種類を採用する。[5]

1. 前にその人に起こっていただろうこと (必要)
2. 後にその人に起こるだろうこと (影響)
3. 前にその人が思っていただろうこと (意図)
4. 後にその人が思うだろうこと (反応)

[5]関係は ATOMIC のそれとまったく同じではない。たとえば本研究における意図は、ATOMIC における xIntent と xWant からなる。



(a) イベント　　(b) 影響の推論

図 5: イベントと推論を獲得するタスクの例

イベント	関係	推論
X が顔を洗う	必要	{ X が水道で水を出す}
	影響	{ X がタオルを準備する, X が鏡に映った自分の顔に覚えのない傷を見つける, X が歯磨きをする}
	意図	{スッキリしたい, 眠いのでしゃきっとしたい}
	反応	{さっぱりして眠気覚ましになる, きれいになる, さっぱりした}

表 10: 小規模な知識グラフの一部

イベントあたり 3 人に尋ね、1 人あたり 1 個の推論を記述してもらう。タスクの例を図 5(b) に示す。重複する三つ組[6]を取り除き、日本語構文解析器 KNP[7]を用いて推論に構文的なフィルタを施す。[8]

クラウドソーシングによって獲得したイベントと推論の統計を表 9 の最左列に示す。コストとしては、計 547 人のクラウドワーカを雇い、その料金は 16,844 円であった。知識グラフの一部を表 10 に示す。

4.1.2　クラウドソーシングによるフィルタ

獲得した推論に対して、それらの品質をクラウドワーカに評価してもらう。推論ごとに適切かどうかを 3 人に判定してもらい、多数決をとる。推論の評価は、関係ごとに独立に行う。

[6]本研究では、あるイベントとそれに対する推論、推論における関係を (イベント, 関係, 推論) という三つ組として扱う。

[7]https://nlp.ist.i.kyoto-u.ac.jp/?KNP

[8]主語が人物 X かどうかや時制が現在かどうか、イベントは 1 文かどうかなどを判定する。

関係	テンプレート
必要	h ためには、t 必要がある。
影響	h。結果として、t。
意図	h のは、t と思ったから。
反応	h と、t と思う。

表 11: ショットのテンプレート

	個数	適切 [%]	Fleiss's κ
イベント	1,471	-	-
必要	9,403	80.81	36.07
影響	8,792	85.45	34.03
意図	10,155	86.06	43.42
反応	10,941	90.30	21.51

表 12: 大規模な知識グラフの統計

適切でないと判定された推論は、フィルタして除去する。

フィルタの統計は表 9 の中 2 列にある。計 465 人のクラウドワーカを雇い、8,679 円を支払った。判定における Inner-Annotator Agreement として計算した Fleiss's κ を、表 9 の最右列に示す。

適切でないと判定された推論には、以下の傾向があった。一つは (X が二度寝する, 反応, 今日は仕事が休みだと思う) のように、前後が逆というものであった。(X がネットサーフィンをする, 必要, 海に着く) のように、自然でないものもあった。

4.1.3　大規模言語モデルによる生成

HyperCLOVA JP Koya 39B モデルを用いて、知識グラフを大規模に拡大する。生成では 10 個のショットを用いて生成を行う。ショットは生成のたびにランダムに選ぶ。

イベント　4.1.1 節で獲得したイベントから、新たなイベントを生成する。ランダムに選んだ 10 個のイベントをショットとして列挙し、モデルに 11 個目を生成させる。10,000 回生成し、重複するイベントを取り除く。プロンプトの例を以下に示す。

1. X がスマホでゲームする
2. X が花に水をやる
3. X が Y を飲み会に誘う
(中略)
11. X が

推論　イベントと同じように、4.1.1 節と 4.1.2 節で獲得した推論をショットとして、10 個の推論から 11 個目を生成する。推論は生成したイベントごとに 10 回生成し、三つ組の単位で重複するものを取り除く。生成では、関係ごとに異なるプロンプトを用いる。ショットの三

イベント	関係	推論
X がコンビニへ行く	必要	{ X が財布を持っている, X が外出する, X が外出着に着替える, X が財布を持って出かける, X が外へ出る, X がジュースを買う, X が財布を持っていく}
	影響	{ X が買い物をする, X が雑誌を立ち読みする, X が ATM でお金をおろす, X が弁当を買う, X がアイスを買う, X が飲み物を買う}
	意図	{何か買いたいものがある, 雑誌を買う, 飲み物を買おう, 飲み物や食べ物を買いたい, なんでもある, 何か買いたい, 朝食を買う, お菓子やジュースを買いたい, 何か飲み物でも買おう}
	反応	{何か買いたいものがある, 何か買う, 何か買おう, 何か買いたくなる, ついでに何か買ってしまう, 何か買ってこよう, 雑誌を立ち読みする, 何も買わない, 便利だ}

表 13: 大規模な知識グラフの例

つ組はテンプレートを用いて自然言語に変換し、パターンマッチによってテールを抽出する。ショットのテンプレートを表 11 に示す。抽出した推論に対して、4.1.1 節の構文的なフィルタを施す。影響の推論を生成するプロンプトの例を以下に示す。

1. X がにわか雨にあう。結果として、X が軒先で雨宿りする。
2. X がネットで服を買う。結果として、X が荷物を受け取る。
3. X が小腹を空かせる。結果として、X が菓子を食べる。
(中略)
11. X が筆箱を忘れる。結果として、X が

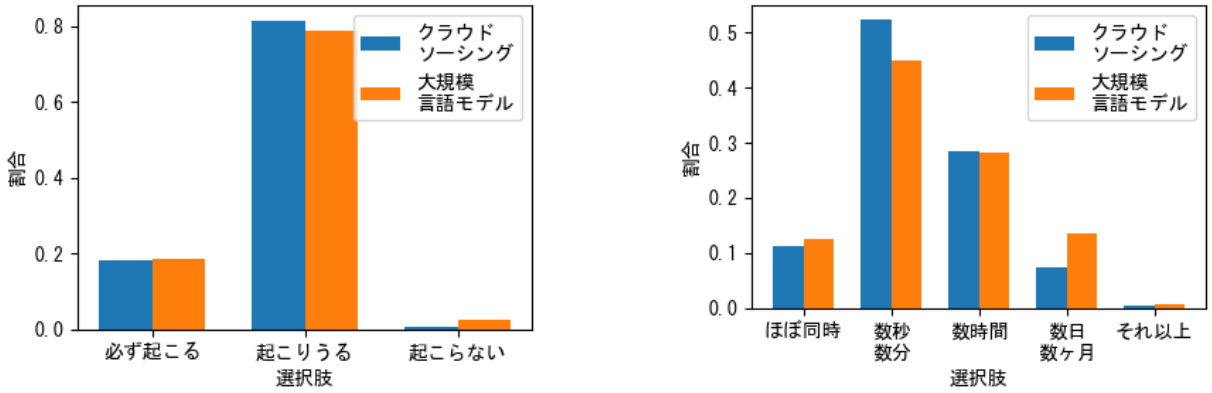
大規模言語モデルが生成したイベントと推論の統計を表 12 の最左列に示す。また生成した推論を 4.1.2 節の手順で評価した結果を表 12 の右 2 列に示す。評価は関係ごとに、生成した推論からランダムに選んだ 500 個に対して行った。計 409 人のクラウドワーカを雇い、その料金は 7,260 円であった。大規模な知識グラフの一部を表 13 に示す。

構築した知識グラフは、たとえば (X が会社に行く, 必要,X が電車に乗る) のように、日本の文化が反映されたものとなっている。このことは、異なる言語の似たような資源をただ翻訳するのではなく、対象の言語においてゼロから構築することの意義を示すとともに、提案する手法の価値を強調している。

4.2　知識グラフの分析

クラウドソーシングの知識グラフと大規模言語モデルの知識グラフ、すなわち人間と言語モデルが生み出す推論の傾向を比較する。本研究では、比較の観点に蓋然性と時間的な幅を採用する。4 関係のうち、代表として影響を検証する。

4.1.1 節と 4.1.2 節で獲得した三つ組のヘッドに対して、4.1.3 節の手順にしたがってテールを 3 個ずつ生成する。影響に関する 554 個の三つ組から、586 個の推論を獲得した。



(a) 蓋然性　　(b) 時間的な幅

図 6: 影響の推論に関する人間と言語モデルの傾向

関係	プロンプト
必要	次の出来事に必要な前提条件は何ですか:
影響	次の出来事の後に起こりうることは何ですか:
意図	次の出来事が起こった動機は何ですか:
反応	次の出来事の後に感じることは何ですか:

表 14: T5 のプロンプト

蓋然性　あるイベントの後に起こるイベントが、どれくらい起こりやすいかを見る。影響の関係にあるイベントのペアをクラウドワーカに与え、後のイベントがどれくらい起こりやすいかを 3 段階で判定してもらう。推論あたり 3 人に尋ね、回答の中央値を採用する。

時間的な幅　あるイベントが起こってから、後に起こるイベントが起こるまでの時間的な幅を見る。蓋然性と同じように、先のイベントが起こってから後のイベントが起こるまでの時間幅を 5 段階で判定してもらう。3 人に尋ね、中央値を採用する。

それぞれの比較を図 6 に示す。図 6(a) から、人間が記述した推論の方がわずかに蓋然的であることがわかる。図 6(b) では、大規模言語モデルが生成する推論の方がより時間的に幅がある。このことから、人間は影響と聞いて比較的すぐ後に起こるイベントを推論するが、言語モデルは少し遠くのイベントを見ると言える。

4.3　常識生成モデルの訓練

4.1 節の知識グラフを用いてイベント常識に関する常識生成モデル [Bosselut 19] を訓練する。大規模言語モデルがもつ知識をより小規模な常識生成モデルに蒸留 [West 22] することで、それらを扱うためのコストが小さくなる。

GPT-2 [Radford 19] と T5 [Raffel 20] の日本語版[9]を、構築した知識グラフで Finetuning する。GPT-2 を用いる場合、関係を表す特殊トークンをイベントの後に付与し、それらを入力

[9]日本語版には、それぞれ `https://huggingface.co/nlp-waseda/gpt2-small-japanese` と `https://huggingface.co/megagonlabs/t5-base-japanese-web` を採用した。

	BLEU	BERTScore	尤もらしさ [%]	スコア
GPT-2	43.61	87.56	92.53	1.79
T5	39.85	82.37	89.58	1.69

表 15: 常識生成モデルの評価

イベント	関係	推論
X がパソコンで仕事をする	必要	X がパソコンを起動する
	影響	X が残業する
	意図	お金を稼ぎたい
	反応	疲れた
X がネットサーフィンをする	必要	X がパソコンを起動する
	影響	X が時間を浪費する
	意図	情報収集したい
	反応	時間が経つのが早い

表 16: GPT-2 に基づく常識生成モデルの生成例

として推論を生成する。一方 T5 では、生成する推論の関係をプロンプトとみなし、入力するイベントの前に記述する。T5 におけるプロンプトを表 14 に示す。

GPT-2 と T5 の日本語版を Finetuning するにあたっては、共通のハイパパラメータを設定する。学習率を 2e-5、Weight Decay を 0.01 とする。Gradient Clipping として、勾配のノルムを最大 1.0 とする。バッチサイズは 16 で、訓練は 3 エポック行う。また生成はすべて Greedy Search で行う。

自動評価として、BLEU [Papineni 02] と BERTScore [Zhang 20b] を計算する。BLEU は、日本語形態素解析システム Juman++ [Tolmachev 18] を用いて分かち書きした単語について計算する。BERTScore は、RoBERTa [Liu 20] の日本語版[10]を用いて計算する。さらに人手評価として、推論の尤もらしさをクラウドソーシングによって評価する。推論の起こりやすさを常に・よく・たまに・決してないの 4 段階で判定してもらう。推論あたり 5 人のクラウドワーカに判定してもらい、決してない以外の判定が過半数を上回る推論を尤もらしいものとする。また上記の 4 段階にそれぞれ 3 から 0 までのスコアを割り当て, 推論ごとに平均をとる。

知識グラフの 9 割を訓練データ、1 割をテストデータとする。テストデータに対する自動評価と人手評価の結果を表 15 に示す。尤もらしさはどちらの常識生成モデルも約 9 割で、おおむね適切な推論を生成していると言える。またすべての指標に関して、GPT-2 が T5 を上回っている。GPT-2 に基づく常識生成モデルの生成例を表 16 に示す。

5 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

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

[10] https://huggingface.co/nlp-waseda/roberta-base-japanese

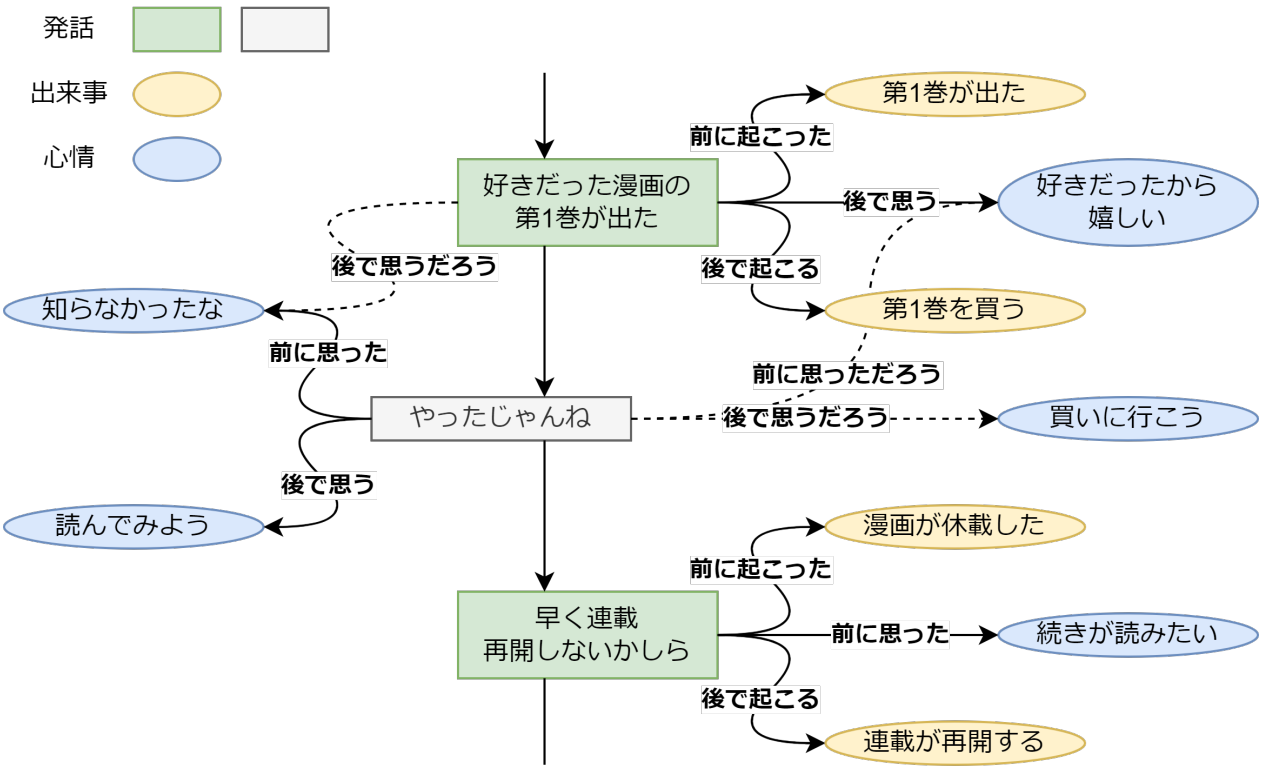
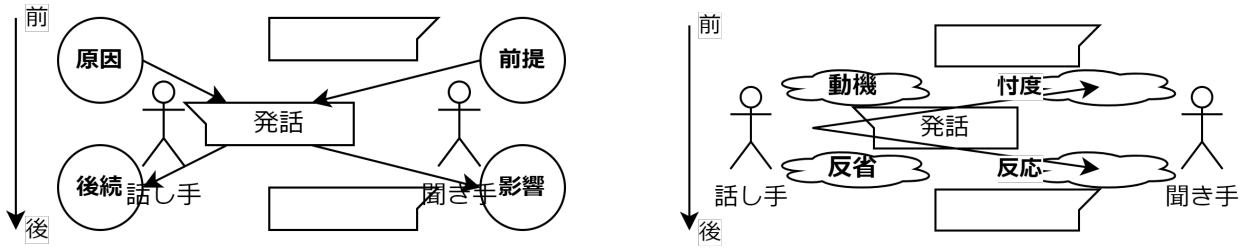


図 7: 発話ごとにタグ付けする推論の例



(a) 出来事　　(b) 心情

図 8: 対話常識グラフで推論すべき関係

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。

5.1　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。

5.1.1　マルチターン対話の収集

Twitter API 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

表 17: 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

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[11]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。

5.1.2　推論の付与

Yahoo!クラウドソーシングを用いて、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

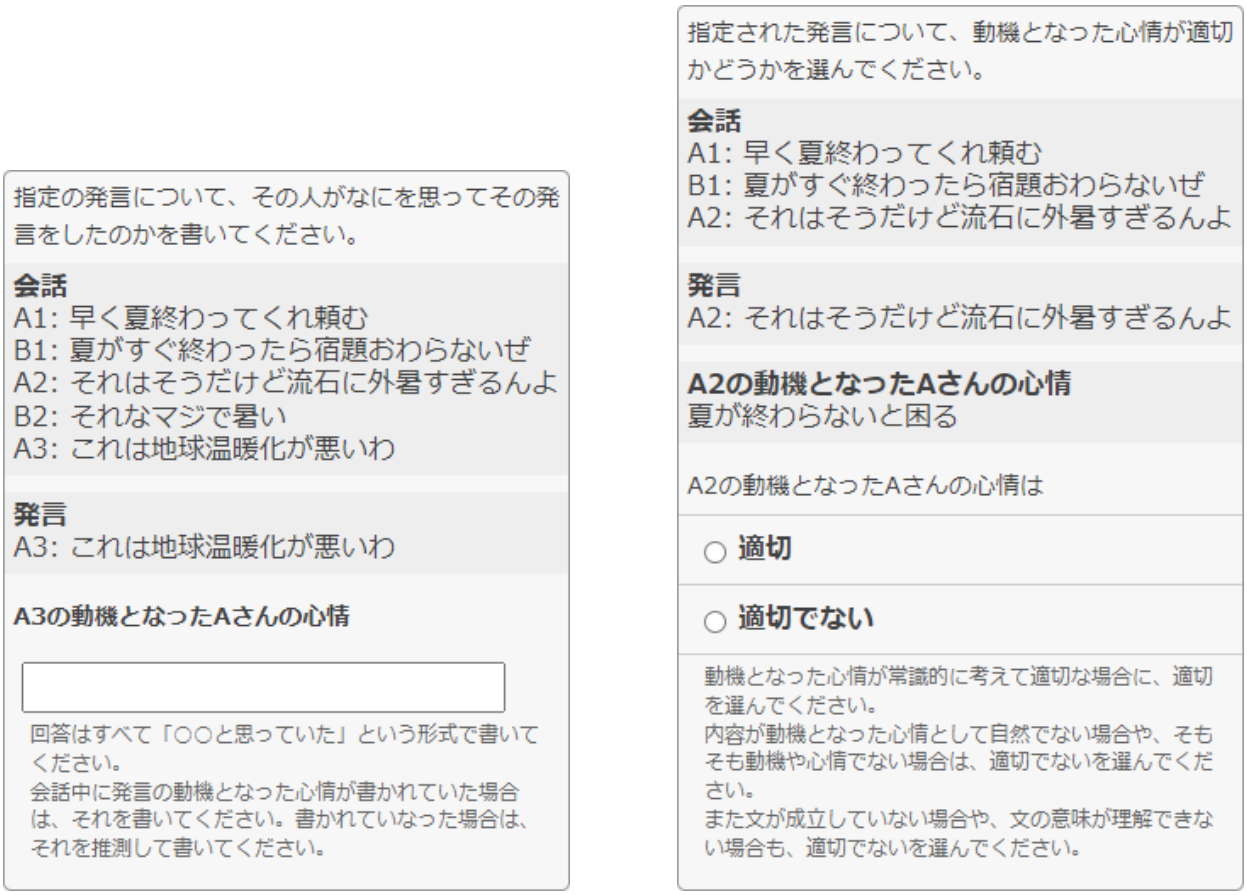
8 関係はそれぞれ、図 8 のように名付ける。例えば (出来事, 前, 話し手) の組合せによる関係は**原因**と呼び、発話の原因となる事象の推論をテキスト形式で付与する。

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 9(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 9(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 17 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。

[11] 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。



(a) 記述　　(b) フィルタ

図 9: 推論を獲得するクラウドソーシングの例

5.1.3　推論の例

5.1.1 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、5.1.2 節で付与した推論を表 18 に示す。

5.1.4　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

表 18: 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}

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。

5.2　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 [Brown 20] 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 [Lee 22] 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[12]。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 [Wei 22] とも言える。

5.2.1　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[13]。各ショッ

[12]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。

[13]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。

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 10: 対話応答生成におけるショットの例

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[14]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 10(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 10(b) に示す。

大規模言語モデルには、HyperCLOVA JP 39B モデル [Kim 21] を用いる。プロンプトのショット数を 2 とする[15]。応答は対話ごとに 3 回ずつ生成する。

[14]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。
[15]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。

表 19: 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

5.2.2　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 19 に示す。

5.2.3　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 [Papineni 02]、distinct [Li 16] を計算する。これらは、日本語形態素解析システム Juman++ [Tolmachev 18] 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[16]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。相対評価の例を図 11 に示す。延べ 763 人のクラウドワーカを雇い、15,350 円を支払った。

[16] https://huggingface.co/nlp-waseda/gpt2-xl-japanese

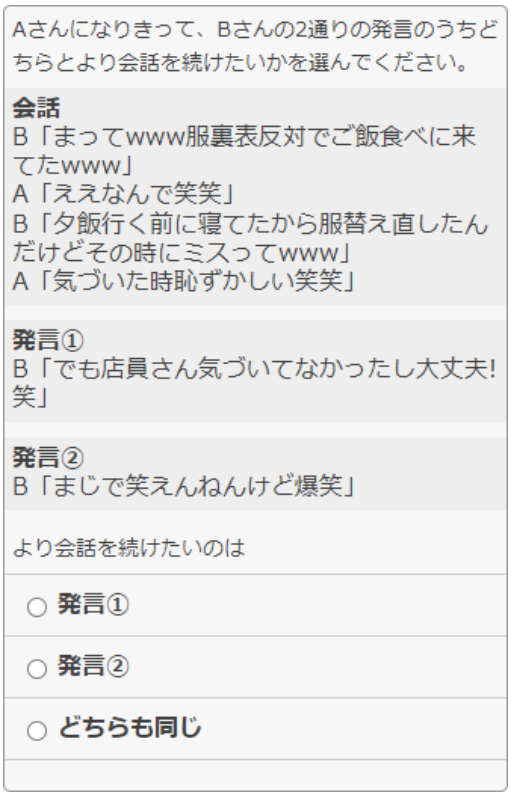


図 11: 相対評価のためのクラウドソーシングの例

表 20: 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

5.2.4　実験結果

自動評価の結果を表 20 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 21 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 20 と表 21 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。

表 21: 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

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。

6 おわりに

人間のように感情や常識を理解する対話システムの開発に向けて、コンピュータの常識理解を促進するデータセットを日本語で構築した。発話の話し手と聞き手が抱く感情をタグ付けした対話コーパスや、イベントの前後で登場人物に起こるイベントや登場人物が思うメンタルステートを収集した常識知識グラフを構築した。またそれらのコンセプトを併せて、発話の前後で話者に起こる出来事や話者が思う心情を収集した常識知識グラフを構築した。それぞれのデータセットを分析した結果、話し手と聞き手が抱く感情の違いや人間が推論する常識の傾向、人間とコンピュータの推論における特性が明らかになった。

Twitter から収集した対話コーパスの品質が低いことや感情タグ付き対話コーパスにおけるタグの分布が偏っていること、常識知識グラフにおける推論に誤りが含まれることなど、問題点が残っている。また常識知識グラフを用いた対話応答生成の実験では、既存の評価手法が対話システムの性能を正しく評価できていない可能性を示した。これらを解決するのは、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。

参考文献

[AlKhamissi 22] AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., and Ghazvininejad, M., “A Review on Language Models as Knowledge Bases” (2022)

[Bosselut 19] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y., “COMET: Commonsense Transformers for Automatic Knowledge Graph Construction”, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pp. 4762–4779, Florence, Italy (2019), Association for Computational Linguistics

[Brown 20] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., "Language Models are Few-Shot Learners", in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. eds., Advances in Neural Information Processing Systems , Vol. 33, pp. 1877–1901, Curran Associates, Inc. (2020)

[Buechel 17] Buechel, S. and Hahn, U., "EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis", in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers , pp. 578–585, Valencia, Spain (2017), Association for Computational Linguistics

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) , pp. 4171–4186, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[Gabriel 21] Gabriel, S., Bhagavatula, C., Shwartz, V., Le Bras, R., Forbes, M., and Choi, Y., "Paragraph-level Commonsense Transformers with Recurrent Memory", Proceedings of the AAAI Conference on Artificial Intelligence , Vol. 35, No. 14, pp. 12857–12865 (2021)

[Ghazi 15] Ghazi, D., Inkpen, D., and Szpakowicz, S., "Detecting Emotion Stimuli in Emotion-Bearing Sentences", in Gelbukh, A. ed., Computational Linguistics and Intelligent Text Processing , pp. 152–165, Cham (2015), Springer International Publishing

[Ghosal 21] Ghosal, D., Hong, P., Shen, S., Majumder, N., Mihalcea, R., and Poria, S., "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 (2021), Association for Computational Linguistics

[Ghosal 22] Ghosal, D., Shen, S., Majumder, N., Mihalcea, R., and Poria, S., "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 (2022), Association for Computational Linguistics

[Hsu 18] Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H., and Ku, L.-W., "EmotionLines: An Emotion Corpus of Multi-Party Conversations", in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) , Miyazaki, Japan (2018), European Language Resources Association (ELRA)

[Hwang 21] Hwang, J. D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., and Choi, Y., "(Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs",

Proceedings of the AAAI Conference on Artificial Intelligence , Vol. 35, No. 7, pp. 6384–6392 (2021)

[Kajiwara 21] Kajiwara, T., Chu, C., Takemura, N., Nakashima, Y., and Nagahara, H., "WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations", in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pp. 2095–2104, Online (2021), Association for Computational Linguistics

[Kim 21] Kim, B., Kim, H., Lee, S.-W., Lee, G., Kwak, D., Dong Hyeon, J., Park, S., Kim, S., Kim, S., Seo, D., Lee, H., Jeong, M., Lee, S., Kim, M., Ko, S. H., Kim, S., Park, T., Kim, J., Kang, S., Ryu, N.-H., Yoo, K. M., Chang, M., Suh, S., In, S., Park, J., Kim, K., Kim, H., Jeong, J., Yeo, Y. G., Ham, D., Park, D., Lee, M. Y., Kang, J., Kang, I., Ha, J.-W., Park, W., and Sung, N., "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 (2021), Association for Computational Linguistics

[Lee 22] Lee, Y.-J., Lim, C.-G., and Choi, H.-J., "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 (2022), International Committee on Computational Linguistics

[Li 16] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B., "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 (2016), Association for Computational Linguistics

[Li 17] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S., "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset", in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) , pp. 986–995, Taipei, Taiwan (2017), Asian Federation of Natural Language Processing

[Liu 17] Liu, V., Banea, C., and Mihalcea, R., "Grounded emotions", in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) , pp. 477–483 (2017)

[Liu 19] Liu, X., He, P., Chen, W., and Gao, J., "Multi-Task Deep Neural Networks for Natural Language Understanding", in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pp. 4487–4496, Florence, Italy (2019), Association for Computational Linguistics

[Liu 20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V., "Ro{BERT}a: A Robustly Optimized {BERT} Pretraining Approach" (2020)

[Mohammad 17] Mohammad, S. and Bravo-Marquez, F., “Emotion Intensities in Tweets”, in Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017) , pp. 65–77, Vancouver, Canada (2017), Association for Computational Linguistics

[Mostafazadeh 20] Mostafazadeh, N., Kalyanpur, A., Moon, L., Buchanan, D., Berkowitz, L., Biran, O., and Chu-Carroll, J., “GLUCOSE: GeneraLized and COntextualized Story Explanations”, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pp. 4569–4586, Online (2020), Association for Computational Linguistics

[Omura 20] Omura, K., Kawahara, D., and Kurohashi, S., “A Method for Building a Commonsense Inference Dataset based on Basic Events”, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pp. 2450–2460, Online (2020), Association for Computational Linguistics

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., “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 (2002), Association for Computational Linguistics

[Paul 92] Paul, E., “An Argument for Basic Emotions”, in Cognition and Emotion , Vol. 6(3/4), pp. 169–200 (1992)

[Petroni 19] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A., “Language Models as Knowledge Bases?”, 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. 2463–2473, Hong Kong, China (2019), Association for Computational Linguistics

[Plutchik 80] Plutchik, R., “A general psychoevolutionary theory of emotion”, in Theories of emotion , pp. 3–33, Elsevier (1980)

[Radford 19] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I., “Language Models are Unsupervised Multitask Learners” (2019)

[Raffel 20] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, Journal of Machine Learning Research , Vol. 21, No. 140, pp. 1–67 (2020)

[Rashkin 18] Rashkin, H., Bosselut, A., Sap, M., Knight, K., and Choi, Y., “Modeling Naive Psychology of Characters in Simple Commonsense Stories”, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 2289–2299, Melbourne, Australia (2018), Association for Computational Linguistics

[Rashkin 19] Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L., “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset”, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pp. 5370–5381, Florence, Italy (2019), Association for Computational Linguistics

[Roemmele 11] Roemmele, M., Bejan, C., and Gordon, A., “Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning.” (2011)

[Sap 19] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y., “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 (2019)

[Sennrich 16] Sennrich, R., Haddow, B., and Birch, A., “Neural Machine Translation of Rare Words with Subword Units”, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 1715–1725, Berlin, Germany (2016), Association for Computational Linguistics

[Shen 22] Shen, S., Ghosal, D., Majumder, N., Lim, H., Mihalcea, R., and Poria, S., “Multiview Contextual Commonsense Inference: A New Dataset and Task” (2022)

[Speer 17] Speer, R., Chin, J., and Havasi, C., “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge”, Proceedings of the AAAI Conference on Artificial Intelligence , Vol. 31, No. 1 (2017)

[Sugiyama 21] Sugiyama, H., Mizukami, M., Arimoto, T., Narimatsu, H., Chiba, Y., Nakajima, H., and Meguro, T., “Empirical Analysis of Training Strategies of Transformer-based Japanese Chitchat Systems” (2021)

[Talmor 19] Talmor, A., Herzig, J., Lourie, N., and Berant, J., “CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge”, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) , pp. 4149–4158, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[Tolmachev 18] Tolmachev, A., Kawahara, D., and Kurohashi, S., “Juman++: A Morphological Analysis Toolkit for Scriptio Continua”, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations , pp. 54–59, Brussels, Belgium (2018), Association for Computational Linguistics

[Wei 22] Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, brian , Xia, F., Chi, E. H., Le, Q. V., and Zhou, D., “Chain of Thought Prompting Elicits Reasoning in Large Language Models”, in Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. eds., Advances in Neural Information Processing Systems (2022)

[West 22] West, P., Bhagavatula, C., Hessel, J., Hwang, J., Jiang, L., Le Bras, R., Lu, X., Welleck, S., and Choi, Y., “Symbolic Knowledge Distillation: from General Language Models to Commonsense Models”, in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pp. 4602–4625, Seattle, United States (2022), Association for Computational Linguistics

[Zellers 18] Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y., “SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference”, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing , pp. 93–104, Brussels, Belgium (2018), Association for Computational Linguistics

[Zhang 19] Zhang, H., Liu, X., Pan, H., Song, Y., and Leung, C. W.-K., “ASER: A Large-scale Eventuality Knowledge Graph” (2019)

[Zhang 20a] Zhang, H., Khashabi, D., Song, Y., and Roth, D., “TransOMCS: From Linguistic Graphs to Commonsense Knowledge”, in Bessiere, C. ed., Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20 , pp. 4004–4010, International Joint Conferences on Artificial Intelligence Organization (2020), Main track

[Zhang 20b] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y., “BERTScore: Evaluating Text Generation with BERT”, in International Conference on Learning Representations (2020)

[柴田 19] 柴田知秀, 河原大輔, 黒橋禎夫, “BERT による日本語構文解析の精度向上”, 言語処理学会 第 25 回年次大会 (2019)

Appendix

A GPT-2 日本語 Pretrained モデル

知識モデルを構築するにあたって、GPT-2 [Radford 19] の日本語版[17]を構築する。日本語版 Wikipedia と CC-100 の日本語部分を訓練データとして、言語モデルを学習させる。

ハイパパラメータは GPT-3 Small [Brown 20] を参考に設定する。学習率を 6e-4、Weight Decay を 0.1 とする。学習率のスケジューラを Cosine とする。バッチサイズは 512 で、訓練は 2 エポック（GPT-3 の約 10%）行う。ステップ数の 0.1%を Warmup に充てる。

B ニューラルネットワークのハイパパラメータ

B.1 大規模言語モデルを用いた推論の生成

4.1.3 節では、生成に関するハイパパラメータを以下のように設定する。生成の最大トークン数を 32 とする。Softmax では 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。

[17]事前学習済みモデルは `https://huggingface.co/nlp-waseda/gpt2-small-japanese` で公開している。

B.2　イベントに関する知識モデルの訓練

4.3 節について、GPT-2 と T5 の日本語版を Finetuning するにあたっては、共通のハイパパラメータを設定する。学習率を 2e-5、Weight Decay を 0.01 とする。Gradient Clipping として、勾配のノルムを最大 1.0 とする。バッチサイズは 16 で、訓練は 3 エポック行う。また生成はすべて Greedy Search で行う。

B.3　対話常識グラフに基づく対話応答生成

5.2.1 節では、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。