# 人間と言語モデルに対するプロンプトを用いたゼロからのイベント常識知識グラフ構築

井手竜也[1] 村田栄樹[1] 堀尾海斗[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.,eiki.1650-2951@toki.,kakakakakakaito@akane.,dkw@}waseda.jp
{takato.yamazaki,shengzhe.li,kenta.shinzato,toshinori.sato}@linecorp.com

## 概要

人間はなにかを理解したり推論したりするとき，常識を用いる．コンピュータにそのような常識を理解させるため，知識ベースを構築する試みがある．しかし，高い品質の大規模なデータを低いコストで獲得するのは難しい．本論文では，クラウドソーシングと大規模言語モデルの両方を用いて，ゼロから知識グラフを構築する手法を提案する．提案手法に従って，翻訳ではない日本語ならではの知識グラフを構築した．さらにその知識グラフから小規模な知識モデルを訓練し，その性能について検証した．

## 1 はじめに

人間はなにかを理解したり推論したりするとき，常識を用いる．それと同じように、コンピュータがオープンドメインの QA を解いたり物語や対話を読んだりするためには，常識が必要となる [1, 2]．コンピュータに常識を理解させるため，それに向けた知識ベースを構築する試みがある．多くはクラウドソーシングによって構築される [3, 4, 5] が，大規模な構築には高いコストがかかる．一方，知識を自動で獲得する手法 [6, 7] もあるが，高い品質の知識を獲得するのは難しい．近年，知識ベースの構築に大規模言語モデル [8] を用いる手法 [9] も提案されている．

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．本論文では，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．

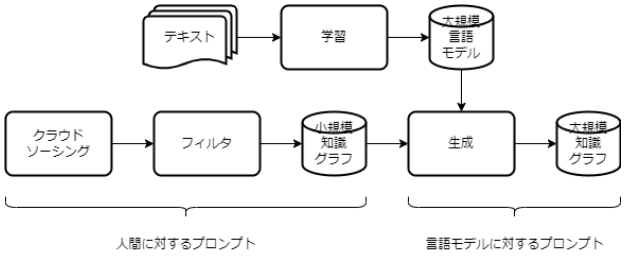



**図 1** 提案手法の概要

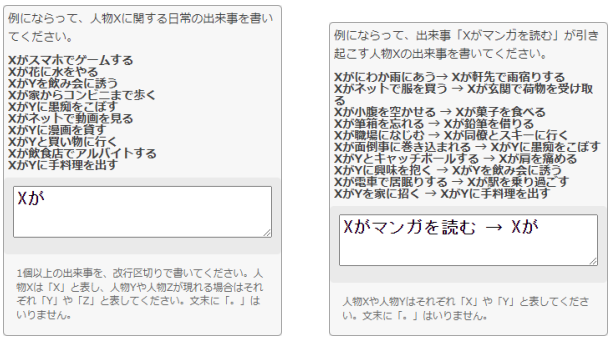
提案手法に従って，イベントに関する常識の知識グラフを構築した．Yahoo!クラウドソーシング[1)] と HyperCLOVA [10] を併せたゼロからの構築によって，英語の単なる翻訳ではない日本語ならではの知識グラフを獲得した．また人間と言語モデルに基づく推論を比較し，傾向の違いを明らかにした．さらに構築した日本語の知識グラフから知識モデル [11] を訓練し，見たことがないイベントに対する推論の生成について検証した．[2)]

## 2 関連研究

ConceptNet [3] はエンティティ同士の関係，ATOMIC [4] はイベント同士もしくは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 [5] がある．また ASER [6] と TransOMCS [7] もともにイベントに関する知識グラフだが，これらは大規模なテキストから抽出されている．

ConceptNet や ATOMIC はクラウドソーシング

1） https://crowdsourcing.yahoo.co.jp/
2） 構築した知識グラフとそれについて訓練した知識モデルは，公開予定である．




(a) イベント　　(b) 影響の推論

**図 2**　イベントと推論を獲得するタスクの例

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．

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 [9] がある．そこでは GPT-3 [8] を用いて ATOMIC を拡大し，RoBERTa [12] を用いてフィルタを施している．

## 3　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 1 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 [4] や ASER [6] 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，二段階目の大規模言語モデルには HyperCLOVA [10] を用いる．

### 3.1　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 2(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．[3)]

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)

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 2(b) に示す．重複する三つ組[4)]を取り除き，日本語構文解析器 KNP[5)] を用いて推論に構文的なフィルタを施す．[6)]

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 1 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 2 に示す．

**表 1**　小規模な知識グラフの統計

| | 個数 | 適切 | 適切 [%] | Fleiss's $\kappa$ |
|---|---|---|---|---|
| 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 |

### 3.2　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．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 1 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 $\kappa$ を，表 1 の最右列に示す．

3)　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．
4)　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．
5)　https://nlp.ist.i.kyoto-u.ac.jp/?KNP
6)　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．

**表 2**　小規模な知識グラフの一部

| 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 3**　ショットのテンプレート

| 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 4**　大規模な知識グラフの統計

| | 個数 | 適切 [%] | Fleiss's $\kappa$ |
|---|---|---|---|
| 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 |

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．

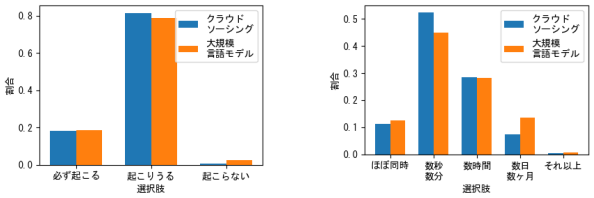
## 3.3　大規模言語モデルによる生成

HyperCLOVA 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**　3.1 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．

**推論**　イベントと同じように，3.1 節と 3.2 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 3 に示す．抽出した推論に対して、3.1 節のフィルタを施す．

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 4 の最左列に示す．また生成した推論を 3.2 節の手順で評価した結果を表 4 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 5 に示す．




(a) 蓋然性　　(b) 時間的な幅

**図 3**　影響の推論に関する人間と言語モデルの傾向

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　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．

3.1 節と 3.2 節で獲得した三つ組のヘッドに対して，3.3 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．

**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．

**表 5**　大規模な知識グラフの例

| 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 6**　知識モデルの評価

| | 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 7**　GPT-2 に基づく知識モデルの生成例

| 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 |

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 3 に示す．図 3(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 3(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．

## 5　知識モデルの訓練

3 節の知識グラフを用いてイベント常識に関する知識モデル [11] 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知識モデルに蒸留 [9] 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 [13] と T5 [14] の日本語版[7]を，構築した知識グラフで Finetuning する．自動評価として，BLEU [15] と BERTScore [16] を計算する．BLEU は，日本語形態素解析器 Juman++[8]を用いて分かち書きした単語について計算する．BERTScore は，RoBERTa [12] の日本語版[9]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 6 に示す．尤もらしさはどちらの知識モデルも約 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 に基づく知識モデルの生成例を表 7 に示す．

## 6　おわりに

本研究では，クラウドソーシングと大規模言語モデルの両方を用いて，ゼロから知識グラフを構築する手法を提案した．提案した手法をもとに，Yahoo! クラウドソーシングと HyperCLOVA [10] を用いて，イベントとメンタルステートの常識に関する知識グラフを日本語で構築した．タスクを設計してクラウドワーカに記述してもらうのと，プロンプトを設計して言語モデルに生成させるのは互いに似ていることから，それらの差異を調査した．さらに構築したイベントの常識に関する知識グラフから，同様の常識に関する知識モデルを訓練した．

ゼロから知識グラフを構築する手法や，それをもとに構築した日本語の知識グラフ，あるいは訓練した知識モデルが，コンピュータの常識推論を促進することを望む．

## 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．

7）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．
8）https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2B
9）https://huggingface.co/nlp-waseda/roberta-base-japanese

# 参考文献


[1] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 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, June 2019. Association for Computational Linguistics.

[2] 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.

[3] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 31, No. 1, Feb. 2017.

[4] 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.

[5] 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. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, No. 7, pp. 6384–6392, May 2021.

[6] Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. Aser: A large-scale eventuality knowledge graph, 2019.

[7] Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. Transomcs: From linguistic graphs to commonsense knowledge. In Christian Bessiere, editor, **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20**, pp. 4004–4010. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[8] 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.

[9] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 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, July 2022. Association for Computational Linguistics.

[10] 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.

[11] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 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, July 2019. Association for Computational Linguistics.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020.

[13] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 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.

[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.

[16] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

# A　生成の詳細

## A.1　プロンプト

イベントの生成では，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する．ショットには番号を振る．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

推論の生成でも，ランダムに選んだ 10 個の推論をショットとして列挙する．ただしショットは自然言語で記述し，パターンマッチによって推論を抽出する．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

## A.2　ハイパパラメータ

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．

# B　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 [13] の日本語版[10)]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 [8] 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．

10）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．

**表 8**　T5 のプロンプト

| 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: |

# C　知識モデルの詳細

## C.1　手法

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 8 に示す．

## C.2　ハイパパラメータ

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．