大規模言語モデルによって構築された常識知識グラフの拡大と低コストフィルタリング

村田栄樹 [1]　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
{eiki.murata.1650-2951@toki.,t-ide@toki.,s.ryoma6317@akane.,dkw@}waseda.jp
{takato.yamazaki,shengzhe.li,kenta.shinzato,toshinori.sato}@linecorp.com

概要

計算機の常識を補うために常識知識グラフが構築されてきたが, 収集コストの観点から規模や言語が限られる. 本研究では, 大規模言語モデルによる常識推論にフィルタモデルを適用することで, 高精度な常識知識グラフの低コストかつ大規模な構築を目指す. まず日本語常識知識グラフにおいて, 生成に関するファクタを分析する. さらに, 常識知識グラフの低コストなフィルタリング手法を提案する. フィルタリングしたグラフを人手評価した結果, 提案手法の有効性が示された. 構築されたグラフで fine-tuning した中規模言語モデルの常識生成に, フィルタリングが与える影響についても検証する.

1　はじめに

計算機による言語理解を実現するためには, 人間がもっているような常識が必要である [1, 2]. そこで, 常識を知識グラフ化する試みがなされてきた. 構築されたグラフは, 常識知識グラフ (CommonSense Knowledge Graph, CSKG) と呼ばれ, その構築は人手によるもの [3, 4, 5] のみならず, GPT-3 などの大規模言語モデル (Large Language Model, LLM) によって行われることもある [6, 7]. また, CSKG で言語モデルを fine-tuning した常識モデルは, LLM よりも高い精度で常識を生成することができる [6, 8].

LLM によって CSKG を構築する場合, 計算機による生成であるため規模を大きくすることができる代わりに, 人手による構築と比較して精度は下がる. そこで, LLM によって構築された CSKG に対して分類モデルによるフィルタリングを行うことで, 規模と精度の双方を高い水準に保つ手法が提案された [6].

しかし, この分類モデルの訓練データは, 生成された CSKG の一部に対して人手アノテーションで作成されるためコストが高い. 一方で, 常識とされる知識は言語や文化圏によって異なることがあり, 単なる翻訳ではなく言語ごとに構築されることが望ましい. 各言語で CSKG を個別に構築する場合, 低コストなフィルタモデルを作成することに需要がある.

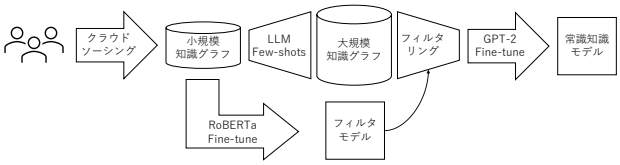


図 1　提案手法の概観.

本論文では, 大規模で高精度な CSKG を低コストで構築することを目的とする. まず日本語 CSKG [7] の拡大のための分析を行い, さらなる拡大が可能であることを示す. 大規模になるに従って適切ではない推論の絶対数も増えると考えられる. そこで, 人手アノテーションを用いない CSKG のフィルタリングの手法を提案する. 既存手法と比較して低コストでありながら有効なフィルタを構築した. また, 構築した CSKG をもとに常識モデルの検証も行う.

2　関連研究

英語における既存の CSKG として, 概念ベースの ConceptNet [3] やイベントベースの ATOMIC [4], それらを統合・拡張した ATOMIC$^{20}_{20}$ [5] がある. これらはすべてクラウドソーシングを用いて構築されている. いずれも 3 つ組で表現され, 例えば ATOMIC は (イベント, 推論の関係, 推論されたイベントやメンタルステート) を要素としている. また, GPT-2 など比較的小さいモデルをこれらの CSKG をもとに fine-tuning することで常識を蓄え, 常識推論をする Transformer [9] として COMET [8] が提案されている.

LLM を用いて CSKG を拡張し, それをもとに

表 1　小規模グラフのサイズを変更したときの生成された大規模グラフのサイズの変化. 1 列目はクラウドソーシングによって収集されたヘッドとなるイベント. 2 列目以降は HyperCLOVA により生成されたイベントや推論.

event (CS)	event	xNeed	xEffect	xIntent	xReact
100	1007	6470	5828	6940	7399
257	1471	9403	8792	10155	10941
392	1429	9182	8442	9893	10497

COMET を訓練する研究もある. West ら [6] は ATOMIC$^{20}_{20}$ を自動拡張した ATOMIC$^{\mathbf{10x}}$ を構築し, 常識モデルとして GPT-2 [10] ベースの COMET$^{\mathrm{DIS}}_{\mathrm{TIL}}$ を訓練した. CSKG の自動拡張には, GPT-3 [11] などに少数の例をショットとして与えてタスクを生成させる few-shot learning の手法が用いられる. さらに, 生成された推論をフィルタリングするためにエンコーダモデルをフィルタとして fine-tuning する. 訓練したフィルタを拡張された CSKG に用いることで, COMET$^{\mathrm{DIS}}_{\mathrm{TIL}}$ は教師である GPT-3 より高精度で常識を生成できる. フィルタの訓練データは LLM により生成された CSKG に含まれる 10,000 個の 3 つ組を人手でアノテーションしている.

日本語における CSKG の構築例もあり [7], 以下の 4 種類の関係からなる CSKG を構築している. また, イベント “X が顔を洗う” に対するそれぞれの推論の例も添える.

- **xNeed** あるイベントの前に人物 X がすること. (例: “X が水道で水を出す”)
- **xEffect** あるイベントの後に人物 X がすること. (例: “X がタオルを準備する”)
- **xIntent** あるイベントの前に人物 X が思うこと. (例: “スッキリしたい”)
- **xReact** あるイベントの後に人物 X が思うこと. (例: “さっぱりした”)

前半の 2 つはイベント対イベントの常識で, 残りの 2 つはイベント対メンタルステートの常識である. クラウドソーシングを用いて小規模にイベントやそれに対する常識推論を収集し, それらをショットとして LLM で CSKG の大規模化を行っている. 日本語の LLM として, HyperCLOVA [12] が使用された. ただし, 英語での研究 [6] のような LLM による拡大後のグラフに対するフィルタリングは行っていない.

3　CSKG の拡大に向けた分析

本節では, 日本語での先行研究 [7] で構築された CSKG に対して追加実験を行い, 日本語 CSKG の分析とさらなる拡大を行う. 大規模グラフを生成する際のパラメータを変更し, HyperCLOVA のもつ知識

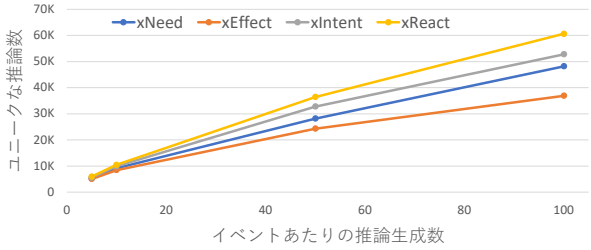


図 2　イベントあたりの推論生成数と生成されたユニークな推論数の関係.

がどの程度 CSKG に移行されるかを検証する. 検証するパラメータは, クラウドソーシングで構築された小規模グラフのサイズと, グラフを拡大する際のイベントあたりの推論の生成回数である. 実験には, HyperCLOVA JP 39B モデルを使用する.

小規模グラフのサイズ　HyperCLOVA の生成のショットとして利用する, クラウドソーシングによって構築される小規模グラフのサイズを変更し, 生成される推論数を比較する. まず, 小規模グラフのイベントをショットに 10,000 回イベントを生成する. さらに, 小規模グラフからランダムに選んだ 3 つ組をショットとして, 生成されたイベントに対する推論を得る (図 1). ショット数および 1 イベントあたりの生成回数をそれぞれ 10 回ずつに固定し, 小規模グラフのイベント数を {100, 257, 392} と変更した場合に生成されたユニークな推論数を比較する.

表 1 に示すように小規模グラフのイベント数が 100 のときは生成されたイベントやユニークな推論数は相対的に小さいが, 257 と 392 のときはおよそ同じ生成数となった. これは, HyperCLOVA の知識を CSKG に引き出すためのショットの多様性として, 小規模グラフのイベント数は 250 から 300 程度で十分であることを示している.

HyperCLOVA による生成回数　先行研究 [7] ではイベントあたりの推論生成回数は 10 で固定していた. これを増やすことで, HyperCLOVA のもつ常識推論の知識を CSKG により引き出すことができると考え, イベントあたりの生成回数を {5, 10, 50, 100} と変化させ生成されるユニークな推論数を比較する. このとき, ショットに用いる小規模グラフやショット数は固定する. 図 2 に示すように, イベントあたりの生成数に対して生成されたユニークな推論数は単調に増加した. 生成回数を増やすことで, 多様な推論を行うことができることがわかる.

まとめると, 一定以上のショットの多様性があれば HyperCLOVA は多様な常識推論をする能力をも

ち, 推論の生成数を増やすことで CSKG のさらなる拡大が実現できる. イベントあたりの推論生成数を 10 から 100 に増やすことでユニークな推論数の合計はおよそ 4 万件から 20 万件に増加した.

4 低コストフィルタリング

ATOMIC^{10x} など [6, 7] では, クラウドソーシングなど人手によって構築された小規模な CSKG をもとに GPT-3 や HyperCLOVA [12] などの LLM を用いて大規模な CSKG を得た. 言語モデルにより生成されるため, 大規模 CSKG の推論には適切ではない推論も含まれる. 3 節で, 生成回数の増加によってさらなる拡大が可能であることがわかったが, 適切ではない推論の割合は同じでもその絶対数は増加する. これまでの CSKG のフィルタモデルは, 人手アノテーションにより訓練データを得ていた [6]. 常識とされるものは言語, 文化により異なるため各言語で CSKG を個別に得ることが望まれるが, 既存の方法のコストは大きい. そこで, 人手アノテーションなしで, LLM により構築される CSKG をフィルタリングする手法を提案する. また, 日本語 CSKG における実験によりその有効性を検証する.

4.1 提案手法

前述の小規模グラフが人手により構築されていることから, 含まれる推論は適切だと仮定する. そこで, 小規模グラフの推論をフィルタモデルの訓練データ (正例) として扱うことで低コストでのフィルタモデルの訓練を試みる. ただし, 分類モデルの訓練には負例も必要であるため, 同じ小規模グラフ内から擬似的に負例を採用する. このように小規模グラフから訓練データを獲得し, 訓練したフィルタで大規模グラフをフィルタリングすることで追加のアノテーションなしで高精度で大規模な CSKG を得る.

以下で, 3 種類の負例の採用方法を提案する. 以降 G_{small} は, 与えられるイベント h (head), 推論のタイプを示す関係 r (relation) と推論 t (tail) の 3 つ組 (h, r, t) を要素とする小規模グラフを表す. 表 2 に訓練データの採用例を示す.

負例タイプ 1 時系列の間違っている負例を擬似的に採用する. 適切な 3 つ組 $(h, r, t) \in G_{\text{small}}$ が与えられたとき, (t, r, h) を考えることで適切ではない推論を得る. ヘッドとテールを入れ替える必要があるため, イベント対イベントの推論のみに用いる.

負例タイプ 2 同じ関係の 2 つの適切な 3 つ組 $(h_1, r, t_1), (h_2, r, t_2) \in G_{\text{small}}, (h_1 \neq h_2)$ から, (h_1, r, t_2) や (h_2, r, t_1) を考えることで適切ではない推論を得る. ヘッドとテールで文脈が異なるため訓練の際は易しい例となり得る.

負例タイプ 3 CSKG には, xIntent と xReact のように時系列的に逆向きの関係が存在する. ある関係 r に対して逆向きの関係を $\text{inv}(r)$ と表すと, 同じヘッドに対する 2 つの適切な 3 つ組 $(h, r, t), (h, \text{inv}(r), t') \in G_{\text{small}}$ に対して, (h, r, t') を負例として採用できる. タイプ 2 とは異なり, 同じ文脈の負例を得ることができる.

4.2 実験

提案手法の検証として, 3 節と同様に日本語 CSKG において実験を行う. 比較として, 先行研究と同様に人手アノテーションによるフィルタモデルも訓練し, 結果を付録 A.2 に示す.

モデル フィルタモデルは事前学習済みのエンコーダモデルを関係ごとに fine-tuning することによって構築する. 訓練はモデルに 3 つ組を入力し, その推論が適切か適切でないかの 2 値分類タスクとして行う. 事前学習済みモデルとして日本語 RoBERTa-large[1)] [13] を用いる. さらに, 先行研究 [6] に倣い, 自然言語推論 (NLI) タスクで追加訓練したモデルを使用する. 日本語の NLI データセットとしては, JGLUE [14] に含まれる JNLI を利用する.

データ 4.1 節に従い, それぞれの関係について小規模グラフに含まれるすべての 3 つ組を正例とする. さらに, 正例の数と同じ数になるように負例を採用する. 負例内の比について, 負例タイプ 2 はやや簡単であると考えたため 2:1:2 とする.

テストデータとして大規模グラフからランダムに抽出し, 人手でラベルを付与した 495 件の 3 つ組を用いる. ラベル付与には Yahoo!クラウドソーシング[2)] を用い, 3 人のクラウドワーカによる多数決により推論が適切か否かを付与した.

結果 訓練したフィルタモデルを用いて, テストデータに対して予測を行う. クラウドソーシングで適切と判断された推論 (1) とそうではない推論 (2) ごとに, 各推論についてモデルが適切と予測した確率の平均を表 3 に示す. すべての関係において, (1) の推論に対する予測確率の方が平均して高くなっていることが確認できる. さらに, 閾値を設けてモデル

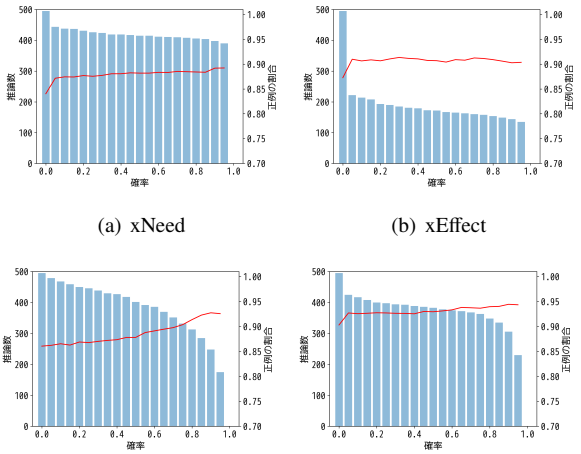
1) https://huggingface.co/nlp-waseda/roberta-large-japanese
2) https://crowdsourcing.yahoo.co.jp/

表 2　小規模グラフから採用する訓練データの例. 各例は (h, r, t) を表す.

負例タイプ	もととなる正例	採用する負例
タイプ 1	(X が顔を洗う, xNeed, X が水を出す)	(X が水を出す, xNeed, X が顔を洗う)
タイプ 2	(X が顔を洗う, xNeed, X が水を出す) (X が PC で仕事をする, xNeed, X が PC を起動する)	(X が顔を洗う, xNeed, X が PC を起動する)
タイプ 3	(X が顔を洗う, xNeed, X が水を出す) (X が顔を洗う, xEffect, X が顔を拭く)	(X が顔を洗う, xNeed, X が顔を拭く)

表 3　クラウドソーシングによって適切だとされた推論とそうでない推論に対して提案手法によるフィルタが適切と予測した確率の平均と標準偏差. 最右列はその差.

関係	適切 (1)	適切でない (2)	(1)−(2)
xNeed	0.880±0.295	0.626±0.450	+0.254
xEffect	0.367±0.447	0.256±0.413	+0.111
xIntent	0.781±0.276	0.643±0.300	+0.138
xReact	0.759±0.348	0.519±0.426	+0.240



(a) xNeed　(b) xEffect

(c) xIntent　(d) xReact

図 3　フィルタモデルによる確率の閾値ごとの通過した推論数 (青, 棒グラフ) と適切な推論の割合 (赤, 折れ線).

による予測確率が閾値以上となる推論のみを採用することでフィルタリングを行う. 図 3 にテストデータに対して閾値ごとのフィルタを通過した推論数とその中の適切な推論の割合を示す. より高い閾値のフィルタをかけることで適切な推論の割合が高くなる傾向を確認できる. よって, 訓練したモデルがフィルタとしての役割をしているといえる.

3 節のようにさらに拡大したグラフに対して, これらのフィルタを適用することで大規模かつ高精度な CSKG を得ることができる.

5　常識モデルの分析

LLM から GPT-2 などの中規模言語モデルへの常識の移行を, 3 節で分析した CSKG の特性と 4 節で構築したフィルタの有無に着目して分析する.

表 4　フィルタリングの強度 s と大規模グラフのイベントあたりの推論生成数 n による常識モデルの精度の変化.

s	0.2	0.5	0.8	1.0
$n = 10$	0.795	0.843	0.855	0.837
$n = 100$	0.893	0.893	0.888	0.893

訓練　日本語 GPT-2-small[3] を CSKG で fine-tuning することで常識モデルを作成する. 訓練データとして, HyperCLOVA によるイベントあたりの推論生成数 $n \in \{10, 100\}$ とフィルタの強度 $s \in \{0.2, 0.5, 0.8, 1.0\}$ によって 8 種類の CSKG を検証する. フィルタは提案手法によるもので行い, CSKG のサイズが元の s 倍になるようにモデルによる確率の低い推論を切り捨てた. $s = 1.0$ の場合はフィルタを適用していないことを表す.

結果　作成した常識モデルにテスト用のイベントを 150 個与え, 生成された推論が適切である割合でモデルの性能を検証する. 推論の評価には Yahoo!クラウドソーシングを使用した. 表 4 に示すように, $n = 10$ の場合にはフィルタの強度が 0.5 や 0.8 のときにフィルタなしのときの精度を上回った. 一方で, $n = 100$ の場合はフィルタリングの強度による常識モデルの精度に差は見られなかった.

元のグラフが小さい場合は, 訓練データ全体に対する適切なものの割合と訓練データのサイズにトレードオフがあり, フィルタの強度を設定することでモデルの訓練にも良い影響を与えたと考えられる. 一方で元のグラフが大きい場合には, フィルタリングを適用しなくとも適切なデータの絶対量が十分に存在し, 差が見られなかったものと考えられる.

6　おわりに

本論文では, 大規模かつ高精度な CSKG を低コストに構築する手法を提案した. まず日本語 CSKG を分析し, 拡大した. さらに, 小規模グラフをもとに低コストなフィルタを訓練し, その有効性を確かめた. また, それらに伴う常識モデルの精度も検証した.

新たな言語や知識グラフの構築の際に提案手法によって高精度化が低コストで行われることを望む.

3)　https://huggingface.co/nlp-waseda/gpt2-small-japanese

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

[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] 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.

[7] 井手竜也, 村田栄樹, 堀尾海斗, 河原大輔, 山崎天, 李聖哲, 新里顕大, 佐藤敏紀. 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. 言語処理学会第 29 回年次大会, 2023.

[8] 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.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

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

[11] 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.

[12] 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.

[13] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In **Proceedings of the 20th Chinese National Conference on Computational Linguistics**, pp. 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China.

[14] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. Jglue: Japanese general language understanding evaluation. In **International Conference on Language Resources and Evaluation**, 2022.

A　フィルタモデル

4 節の詳細を述べる.

A.1　訓練の詳細

提案手法のフィルタモデルの訓練の詳細を記す. 小規模グラフから採用した訓練データのサイズを表 5 に示す.

“[CLS] head [SEP] tail [SEP]” を RoBERTa に入力し, fine-tuning する. 実験は全て JNLI で追加訓練済みのモデルを利用する. JNLI は 3 値分類タスクであるため, RoBERTa のエンコーダ部分のみを利用し, 2 値の分類ヘッドを新たに追加して行う.

学習は 20 エポック行い, F1 値が最も高いチェックポイントをモデルとして採用した. また関係ごとに, 学習率, バッチサイズとウェイトディケイについてパラメータ探索を行った.

A.2　人手アノテーションによるフィルタ

比較のために, 既存の手法によるフィルタも訓練する. フィルタリング対象である大規模グラフからランダムに推論を選び, 人手によって適切か否かのラベルを付与する. そのラベルをもとに RoBERTa を fine-tuning する. モデルやハイパーパラメータは提案手法によるものと同様である.

提案手法の表 3 と図 3 と同様のものを, それぞれ表 6 と図 4 に示す. 図 4 のように確率の閾値が高くなるしたがって適切な推論の割合が高くなっていることは確認できるが, 表 6 では提案手法ほどの差は見られない. 提案手法と異なり, 訓練データの正負比が偏るため予測確率も 1 に偏ったことが考えられる.

表 5　提案手法によるフィルタの訓練データ数. 関係ごとに, 正例と負例を合わせた数字を示す.

	xNeed	xEffect	xIntent	xReact
推論数	1,401	1,750	1,730	1,861

表 6　クラウドソーシングによって適切だとされた推論とそうでない推論に対して, 人手アノテーションによるフィルタモデルが適切と予測した確率の平均と標準偏差. 最右列はそれらの差.

関係	適切 (1)	適切でない (2)	(1)-(2)
xNeed	0.819±0.017	0.813±0.022	+0.006
xEffect	0.926±0.022	0.913±0.030	+0.013
xIntent	0.936±0.227	0.694±0.434	+0.242
xReact	0.919±0.017	0.917±0.020	+0.002

A.3　取り除かれた例

例として提案手法のフィルタによって適切である確率が 5%未満とされた, CSKG の 3 つ組を示す.

- Xが晩御飯を作っている **xNeed** Xが夕食を食べる
- XがYとドライブする **xEffect** Xが運転免許を取得する
- XがＴＶを見ながらお菓子を食べる **xIntent** おいしい
- Xが銀行のＡＴＭで現金を引き出す **xReact** お金が欲しい

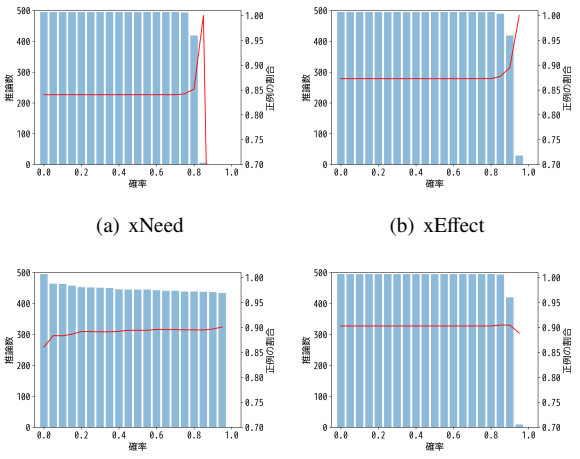
B　常識モデル

5 節の詳細を記す.

訓練は, 3 つ組 (h, r, t) を連結したものをデータとして GPT-2 を fine-tuning することで行う. xNeed など関係はモデルの語彙に存在しないため, スペシャルトークンとして追加しトークナイズする. どの CSKG に対しても 3 エポックの学習を行う.

テストの際は, 3 つ組のうちヘッドイベントと関係を入力し, 続きを生成させることで推論（テール）を得る. 生成時のビームサーチの本数は 5 とする.

また, 表 7 に既存手法と提案手法のフィルタの比較を示す.



(a) xNeed　(b) xEffect　(c) xIntent　(d) xReact

図 4　人手アノテーションフィルタモデルによる確率の閾値ごとの推論数 (青, 棒グラフ) と適切な推論の割合 (赤, 折れ線).

表 7　推論生成回数 n の異なる CSKG に各種フィルタを適用し, 常識モデルを訓練したときの精度の比較. 既存手法は付録 A.2 で訓練したもの. フィルタ強度は $s = 0.8$ で固定した.

フィルタ	$n = 10$	$n = 100$
提案手法	**0.855**	0.888
既存手法	0.853	0.867
なし	0.837	**0.893**