事前学習コーパス内の特定の属性への言及の急激な変化の調査

大萩雅也, 綿岡晃輝, 高山隼矢, 吉川克正
SB Intuitions 株式会社
{masaya.ohagi, koki.wataoka, junya.takayama, katsumasa.yoshikawa}@sbintuitions.co.jp

概要

Web からクロールされた事前学習コーパス内の有害性を含む文章が LLM の有害な出力を引き出すことは知られているが、事前学習コーパス内の有害性が時とともにどのように変化するかは十分に調査されていない。特に、紛争の勃発や移民問題などの社会問題を通じて 2,3 年単位で世論に起きた急激な変化がどのようにコーパスに影響を与えるかは未だ調査されていない。本研究では近年その取り巻く環境が大きく変わった民族に対する言及がどのように変化しているかを、事前学習に用いられることの多い CommonCrawl の日本語データを対象として調査する。結果として CommonCrawl では紛争などの事象が発生した時期を境として、特定の属性に対する有害な文脈を伴う言及が増加することが確認された。

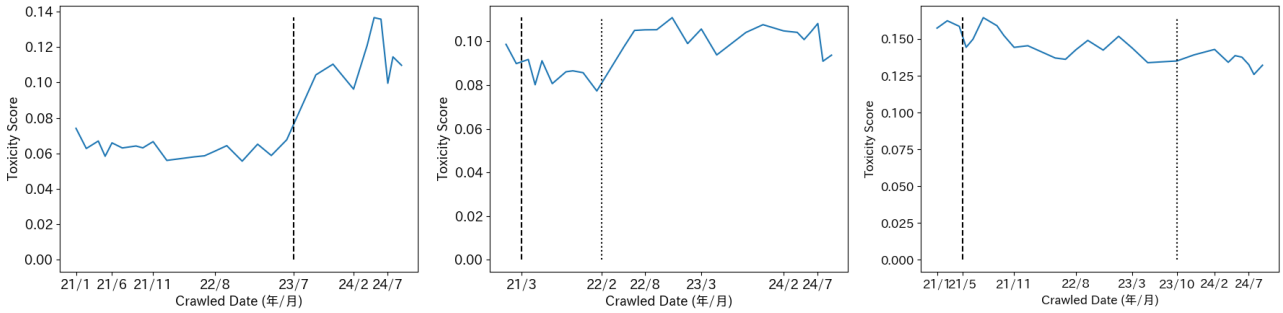
1 はじめに

GPT-3 [1] や Llama [2] に代表される大規模言語モデル (LLM) はその事前学習の過程において Web からクロールされた大規模コーパスがよく用いられる。コーパスにはマイノリティや特定の民族に対する攻撃的な表現がしばしば含まれており、LLM は事前学習時にそのような表現を出力するように学習してしまう [3]。LLM が出力する有害表現に関する既存研究は主にコーパスと LLM の出力の間の関係の調査や、また LLM の有害な出力を抑えるための手法の開発に取り組んできた [4, 5]。しかしながら、ほとんどの既存研究は特定のタイムスタンプのコーパスのみを対象にしており、複数のタイムスタンプのコーパス間にどのような違いがあるかの調査は行なっていない。その時々の社会の時流によって Web に書き込まれる内容は変わり、それはクロールされたコーパスにも影響することが考えられる。社会の時流がどのようにコーパスに影響しているのかを調査するのは LLM が今後長く使用されることが期待される今の時代において重要な課題である。

我々と近いモチベーションをもつ研究としては Yi らのもの [6] が挙げられる。この研究は 2020 年から 2022 年の各月の SNS データを対象とし、それぞれのタイムスタンプのコーパスで言語モデルを訓練することにより性別バイアスや年齢バイアスが時系列順にどのように変化するかを調査した。この研究では一部を除いたほとんどのタイプのバイアスにおいて時系列による変化が見られなかったという実験結果が得られている。しかしながら、そもそも性別バイアスや年齢バイアスのような社会に深く根差したバイアスは数年で大きく変わるものではない。時系列による LLM の社会バイアスの変化をより深く知るためには、近年急激に社会の中でその立ち位置が変わりつつある属性に目をむける必要がある。そこで本研究はここ数年で大きく取り巻く環境が変わった民族にフォーカスをあて、それぞれの民族に対する言及とその有害性がコーパス内でどのように変化しているかの調査を行う。

本稿ではクルド人、ロシア人、ユダヤ人の 3 つの民族を対象属性、LLM の事前学習においてよく用いられる CommonCrawl の 2021 年から 2024 年までの日本語データを対象データとして定量的な分析と定性的な分析を行う。定量的な分析としては各属性に対する言及の有害性がそれぞれの時期ごとにどのように変化し、またその言及の量がどのように増減しているかを測定する。定性的な分析としては属性に対応する文字列周辺に共起する名詞の出現頻度を測り、それを時期ごとに比較することでどのような文脈で属性への言及が発生しているかを分析する。

分析の結果、図 1 に示すようにクルド人とロシア人に対しては特定の時期を境目として、その言及の有害性の急激な増加が確認された。これらは紛争の勃発など実社会の事象と強い関連を見せている。特にクルド人に関する言及の有害性はその量とともに 2023 年以降急激な増加を見せており、社会問題に対応して事前学習コーパス内に今までは存在しなかった新たな有害性が発生しうることが確認された。



(a) クルド人　(b) ロシア人　(c) ユダヤ人

図 1: 各属性に対応する文字列が含まれている文章の有害性の時系列遷移。縦軸の数値は Perspective API によりアノテーションした有害性のスコア。黒の縦線はそれぞれ川口市立医療センターでの騒動が起きた 2023 年 7 月, ロシアとウクライナの間の緊張が高まった 2021 年 3 月, ロシアがウクライナに侵攻した 2022 年 2 月, イスラエルとガザの紛争が起きた 2021 年 5 月, 2023 年 10 月を示している。

2 実験

本研究では事前学習コーパス内での特定の属性に対する言及やその有害性が時が経つごとにどのように変化するかを調査する。本章ではその対象属性、対象コーパス、そして調査手法について説明する。

2.1 対象属性

調査を行う対象の属性として、クルド人、ロシア人、ユダヤ人の 3 つの民族を選定した。それぞれについて選定理由を記述する。

クルド人は特に埼玉県川口市における住民との軋轢が 2023 年以降国内で注目を集めている [7, 8, 9]。この背景には 2023 年以降のクルド難民の申請急増があると言われており、軋轢の表面化の一例として 2023 年 7 月の川口市立医療センターでの騒動が挙げられる [10]。次に、ロシア人に関しては 2022 年 2 月のウクライナ侵攻をきっかけとして対露感情が悪化していることが挙げられる。2023 年度の日本国内の世論調査における「ロシアに対して親しみを感じる」と答えた人は過去最低の割合を記録しており [11]、ロシア国民に対する感情にも一定の影響があると考えられる。ロシア人と同様に、ユダヤ人もイスラエルによる 2023 年の 10 月のガザ侵攻という紛争を抱えている。以上の 3 つの民族はここ数年で取り巻く環境が急激に変わった属性であり、本研究の対象として適したものであると言える。

なお、本稿はこれらの民族と関連する紛争や移民問題に対する是非を判断する立場にはなく、着目するのはあくまでそれぞれの民族を一括りにした差別的な発言であったり攻撃性を含んだ過激かつ有害な発言であることに注意されたい。

2.2 対象コーパス

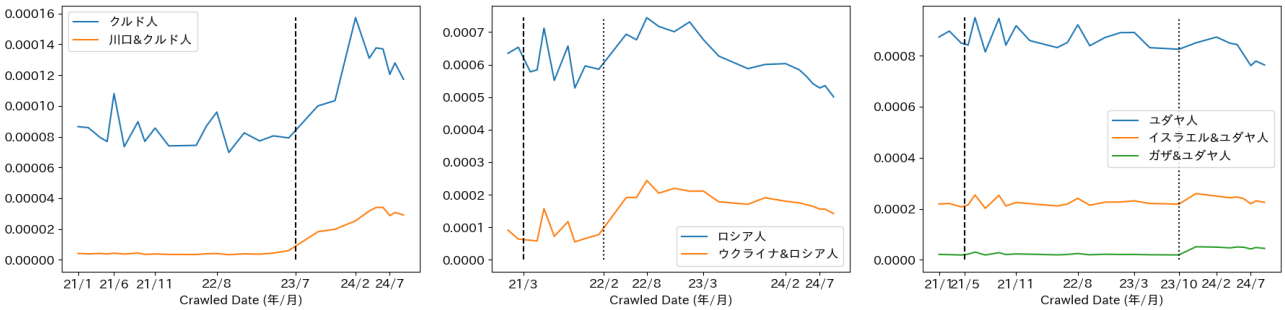
2021 年の第 4 週から 2024 年の第 38 週までにクロールされた CommonCrawl[1] を解析対象のデータとする。CommonCrawl は 1 週間から 2 週間に一度のペースでクロールされたデータを公開している。例えば CC-MAIN-2023-50 はその前のバージョンである CC-MAIN-2023-40 が dump された 2023 年 40 週から 2023 年 50 週までの 10 週間でクロールされたデータとなっている。このデータをタイムスタンプ CC-MAIN-2023-50 のクロールデータ、データ内のドキュメント数をクロール数と呼ぶ。合計 27 個のタイムスタンプそれぞれのクロールデータに対して分析を行い、それを時系列で並べることで事前学習コーパスにおける変化を調査する。

なお、CommonCrawl の生データは重複や非文を多く含んでいる。我々は CCNet [12] を用いて重複ドキュメントを削除し、また本研究が日本国内の世論を対象としたものであるため日本語のドキュメントのみを抽出した。その上で HojiChar[2] を用いて非文や繰り返し文の除去を行なった。

2.3 調査手法

本稿では対象データに対し定量的な解析と定性的な解析を行う。まず定量的な解析の一つ目として、クロールデータの中の対象属性への言及の有害性

1) https://commoncrawl.org/
2) https://github.com/HojiChar/HojiChar



(a) クルド人　(b) ロシア人　(c) ユダヤ人

図 2: 全クロールドキュメント数のうち各属性に対応する文字列が含まれているドキュメントの割合の時系列遷移。縦線はそれぞれ川口市立医療センターでの騒動, ウクライナとロシアの間の緊張が高まった二つの時期, イスラエルとガザの紛争が起きた二つの時期を示している。

が時系列に沿ってどのように変化しているかを調査する。具体的には、各対象属性に対応する文字列 (クルド人, ロシア人, ユダヤ人) が含まれている文章に対して Perspective API [13, 14] を用いて、0.0(無害) から 1.0(有害) までのスコアを自動でアノテーションし、各タイムスタンプごとにその平均を取る。なお、Perspective API の rate limit によりすべての文章を分析するのは困難であるため、クルド人に対しては全体の 10%, ロシア人とユダヤ人に対しては 5%をサンプルして分析を行なった。

また、各属性に対応する文字列が含まれているドキュメント数が時系列にそってどのように増減しているかを調査する。具体的にはクロール数に対しての対象文字列を含むドキュメント数の割合を各タイムスタンプごとに計算する。

定性的な解析としては、対象属性に対応する文字列の前後の 20 単語ずつ、合わせて 40 単語の中での名詞の出現頻度を調査する。頻出する単語がどう移り変わっていくかを分析することでそれぞれのタイムスタンプごとにどういう文脈で対象属性が言及されているのかを比較、調査する。

3 結果

3.1 定量的な解析

各属性に対応する文字列を含む文章の有害性の推移が図 1 に示されている。また各属性を含むドキュメント数の割合の遷移が図 2 に示されている。

クルド人に関しては 2023 年以前と以後で大きな違いが見て取れる (図 1(a), 2(a))。具体的には、2023 年 7 月以降で有害性、ドキュメント数ともに大きな増加を見せている。これは川口市立医療センターでの騒動が起きた時期と一致しており、ここからは実社会で起きた事象の影響がすぐさまクロールデータに現れることが見て取れる。川口とクルド人を両方含むドキュメントの出現割合に関して見てみると ((図 2(a)) その傾向がより顕著に現れており、2023 年 7 月以前はほぼ 0%に近い出現の割合であったものがそれ以降はクルド人を含むドキュメント数の推移と連動しながら強い増加傾向を見せている。

なお、クルド人難民問題自体は日本のみならず全世界的な問題であるため、2023 年以前もクルド人に言及するドキュメント数は一定数存在した。しかしながら 2023 年 7 月以降今までにない激しい増加を見せていることから、2023 年まではあくまで国際的な問題の一つであったクルド人問題が日本国内の問題として捉え直されていった過程が見て取れる。ここについては定性的な解析で詳しく論じる。

ロシア人に関しても同様にウクライナ侵攻をきっかけとして有害性, ドキュメント数が上昇している (図 1(b), 2(b))。ただし、ロシア人に関しては 2022 年 2 月のウクライナ侵攻以前にも有害性、ドキュメント数の増減が見られる。これは 2021 年 3 月から 4 月にかけてロシアとウクライナの間で緊張が高まったこと [15] が関連していると考えられる。しかしながら現在続いている紛争のような長期的な事象には発展しなかったこともあり、有害性、ドキュメント数ともに一時的な増加の後は減少している。

ユダヤ人に関しても 2021 年、2023 年のガザ紛争を契機として有害性、ドキュメント数の増加が見られる (図 1(c), 図 2(c))。しかし、クルド人やロシア人に比べると上昇幅は小さく、すぐに紛争以前の状態

に戻っている。特に有害性に関しては全体として減少傾向を見せている。注目すべき点としては紛争以後ガザとユダヤ人が共起しているドキュメント数は高止まりを続けていることが挙げられる (図 2(c))。ここからは、ガザ問題とユダヤ人を紐付けた言及自体は増えているが、全体としてドキュメント数や有害性に影響を与えるほどではなかったということがわかる。3.2 節で後述するようにユダヤ人はガザ問題以外にも語られる文脈が多いためだと思われる。

以上の分析をまとめると、紛争や移民問題といった実社会の事象は、その事象と結びつきが強い属性に対する Web 上の言及、そしてそのクロールデータに大きな影響を与えることがわかる。その影響は有害性だけでなくその言及の量の部分においても観察され事前学習による LLM への伝搬が危惧される。なお、クロールデータという性質上事象の発生から一定の期間を空けてデータに反映されるのではないかと当初予想していたが、それに反し有害性、量ともにすぐにデータへの影響が現れていた。

3.2 定性的なデータ解析

表 1 に 2021 年 4 月と 2024 年 7 月のクルド人との共起語を出現頻度順に上位 5 つずつ載せている。どちらにおいてもクルド民族と結びつきが強いトルコが最も出現頻度が高いが 2 位以降に差が現れている。2021 年ではシリア内戦やイラク内のイスラム国掃討 [16, 17] に関連した文脈で語られることが多く出現頻度もそれを反映しているが、2024 年の出現頻度では日本や難民問題という単語の出現頻度がかなり上昇している。「川口市」や「在日」といった日本国内の問題に関連した単語も 2024 年時点では上位 50 位圏内まで上昇しており、ここからは単にクルド人の属性に対する言及の量や有害性が上がってきているだけではなく、その語られる文脈が大きく変わってきていることが見て取れる。また、2023 年以降の特徴としては「名無し」や「ハムスター」といったいわゆるまとめサイトと関連する単語の出現頻度が上がってきていることが挙げられる[3)]。クルド人問題が国際ニュースではなく日本国内の社会問題としてネット掲示板などで取り沙汰されるようになってきていることの証左であると言えるだろう。

ロシア人に関しても 2021 年と 2024 年では紛争の影響により違いが見られる (表 2)。2024 年時点ではウクライナやその侵攻を指示したプーチンに関する言及が増加している。一方、ユダヤ人に関しては上位 5 件に大きな変化は見られない (表 3)。ユダヤ人はキリスト教、ホロコーストなどパレスチナ問題以外にも登場する文脈の種類が幅広いため、パレスチナの共起頻度は上がっているものの上位 5 件に強く影響を与えるほどではなかったと思われる。3.1 節でもユダヤ人に言及するドキュメントは有害性、量ともに他の二つの民族ほどの大きな変化は起こっておらず、それと一致する結果となった。

	1	2	3	4	5
21/4	トルコ	シリア	イラク	日本	勢力
24/7	トルコ	日本	難民	シリア	問題

表 1: 2021 年 4 月と 2024 年 7 月のクルド人と共起する名詞の出現頻度の上位 5 語

	1	2	3	4	5
21/4	ロシア	日本	女性	名前	ロシア語
24/7	ロシア	日本	ウクライナ	女性	プーチン

表 2: 2021 年 4 月と 2024 年 7 月のロシア人と共起する名詞の出現頻度の上位 5 語

	1	2	3	4	5
21/4	イスラエル	ドイツ	日本	イエス	世界
24/7	神	イエス	イスラエル	ドイツ	世界

表 3: 2021 年 4 月と 2024 年 7 月のユダヤ人と共起する名詞の出現頻度の上位 5 語

4 おわりに

本研究では CommonCrawl を対象として大規模 Web コーパス内の特定の民族への言及が時系列ごとにどのように変化していくかを調査した。結果として、実社会の事象と連動してクルド人とロシア人への言及が有害性と量の両方で急激に増加していることが確認された。今後はこれらのコーパスのタイムスタンプごとに LLM を訓練して事前学習コーパス内の急激な変化がどのように LLM に伝搬していくかを調査する。また、クルド人のようにある時期を境として有害性が急激に上昇するような属性をどのように検知していくかも重要なテーマである。これらはデータ内の有害性を自動で検知する仕組みとデータの背後にある社会問題に対する深い知識や理解の両方が必要であるため、それらを組み合わせた学際的なアプローチが重要であると思われる。

3) 具体的にはハムスター速報 (https://hamusoku.com/) や 5ch (https://itest.5ch.net/) に関連している

参考文献

[1] 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.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2405.14734**, 2024.

[3] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics.

[4] Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Adding instructions during pretraining: Effective way of controlling toxicity in language models. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 2636–2651, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[5] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 6032–6048, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Yi Zhou, Danushka Bollegala, and Jose Camacho-Collados. Evaluating short-term temporal fluctuations of social biases in social media data and masked language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 19693–19708, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] NHK 首都圏ナビ. 埼玉川口市がクルド人めぐり国に異例の訴えなぜ現場で何が. https://www.nhk.or.jp/shutoken/wr/20240202a.html[アクセス日: 2024 年 12 月 28 日], 2024 年 2 月 2 日.

[8] 産経新聞. 川口のクルド人はなぜ増えたかきっかけはイラン人、民主党政権で難民申請激増. https://www.sankei.com/article/20240502-5QEKJJWHPJPCBLXBZ3XQYKXNBQ/[アクセス日: 2024 年 12 月 28 日], 2024 年 5 月 2 日.

[9] 朝日新聞. クルド人団体事務所周辺のデモ さいたま地裁が実施禁じる仮処分決定. https://www.asahi.com/articles/ASSCP0RR7SCPUTNB013M.html[アクセス日: 2024 年 12 月 28 日], 2024 年 11 月 21 日.

[10] 埼玉新聞. 乱闘…男女もめた結果、病院で 100 人大騒ぎ 救急車受け入れできず 批判の先は…誤解も騒動を取材＜上＞. https://www.saitama-np.co.jp/articles/40494/postDetail[アクセス日: 2024 年 12 月 28 日], 2023 年 8 月 12 日.

[11] NHK. 「親しみ感じる」中国 12.7 % ロシア 4.1 % 過去最低 内閣府調査. https://www3.nhk.or.jp/news/html/20240119/k10014327041000.html[アクセス日: 2024 年 12 月 28 日], 2024 年 1 月 19 日.

[12] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[13] 小林滉河, 水本智也, 佐藤敏紀, 浅原正幸. Japanese real toxicity prompts: 日本語大規模言語モデルの有害性調査. Technical Report 29, SB Intuitions 株式会社, 国立国語研究所, nov 2023.

[14] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In **Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, KDD '22, p. 3197–3207, New York, NY, USA, 2022. Association for Computing Machinery.

[15] 防衛研究所. 2021 年春のウクライナにおけるエスカレーション危機. https://www.nids.mod.go.jp/publication/commentary/pdf/commentary165.pdf [アクセス日：2024 年 12 月 28 日], 2021 年 5 月 13 日.

[16] アジア経済研究所. シリア内戦とクルド民族主義勢力. https://www.ide.go.jp/library/Japanese/Publish/Reports/AjikenPolicyBrief/pdf/084.pdf [アクセス日：2024 年 12 月 28 日], 2017 年 3 月 31 日.

[17] アジア経済研究所. 「イスラーム国（is）」後のクルド問題――困難に直面するクルド人勢力. https://www.ide.go.jp/library/Japanese/Publish/Reports/AjikenPolicyBrief/pdf/128.pdf [アクセス日：2024 年 12 月 28 日], 2019 年 5 月 13 日.

A 事例分析

有害な表現を含むため，論文PDFと分離して公開します．
下記URLから参照してください．
※全角になっていますのでご注意ください

ｈｔｔｐｓ：／／ｗｗｗ．ａｎｌｐ．ｊｐ／ｎｌｐ２０２５／ｐｄｆ／４３０ａ．ｐｄｆ

上記URLにアクセスできない場合は，著者に直接お問い合わせください．