汎用言語モデルは日本語の助数辞を理解しているか

小谷野華那 [1]　谷中瞳 [2]　峯島宏次 [3]　戸次大介 [1]
[1] お茶の水女子大学　[2] 東京大学　[3] 慶應義塾大学
{koyano.kana,bekki}@is.ocha.ac.jp　hyanaka@is.s.u-tokyo.ac.jp
minesima@abelard.flet.keio.ac.jp

概要

日本語には，様々な数量表現の出現形式や助数辞があり，言語学の理論的研究の対象となっている．英語や日本語で，言語学の分析に基づいた数量表現コーパス，推論データセットが構築され，汎用言語モデルが数量表現の理解を必要とする推論をどの程度扱えるかの分析が進められている．しかし，現在の汎用言語モデルには，数量表現の推論の扱いに課題がある可能性が示唆されている．本研究では，その原因の一つとして，汎用言語モデルの事前学習時に，数量表現の振る舞いが正しく学習されないという仮説を立て，助数辞がマスクされたテストセットを構築し，汎用言語モデルが助数辞を正しく予測するかを調査する．実験の結果，予測結果はテスト文に含まれる他の数量表現の影響を受けており，汎用言語モデルは，数量表現の意味を正しく捉えられていない可能性があることを示した．

1　はじめに

日本語には様々な数量表現の出現形式や助数辞があり，言語学の理論的な研究が進んでいる [1, 2, 3, 4, 5]．さらに，日本語の助数辞の違いは，推論に影響を与える．次の例の (1) と（2) は，それぞれ「人」「名」という助数辞を含んでいるが，この 2 つの助数辞はどちらも人数を表す助数辞であるため，(1) と（2) の間には含意関係が成立する．一方，(3）は，「箱」という助数辞を含んでおり，この助数辞は人数を表す助数辞ではないため，(1) と（3）の間では含意関係が成立しない．

(1)　ペットボトルが 3 人分ある．
(2)　ペットボトルが 3 名分ある．
(3)　ペットボトルが 3 箱分ある．

Koyano et al. [6] では，言語学における日本語の数量表現に関する理論を整理し，日本語の実テキストに含まれる数量表現にアノテーションを行い，数量表現コーパスを構築した．さらに，数量表現コーパスから数量表現に関する自然言語推論データセットを構築し，汎用言語モデルが数量表現の理解を必要とする推論をどの程度予測できるかを調査する実験と分析を行い，現在の汎用言語モデルは数量表現の推論に課題を残していることを示唆した．

本研究では，汎用言語モデルにおけて数量表現を含む推論に課題がある原因として，事前学習時に数量表現の振る舞いが正しく学習されていないという仮説を立てる．この仮説を検証するために，Koyano et al. [6] の数量表現コーパスを用いて，助数辞をマスクしたテストセットを構築する．このテストセットを用いて，マスクした箇所に入る助数辞を正しく予測しているかどうかを調査する．汎用言語モデルが正しい助数辞を予測できていれば，助数辞の前後の文脈は正しく捉えられていると考えられる．そして，実験結果から，汎用言語モデルが日本語の数量表現の助数辞をどの程度捉えられているかを分析する．

2　日本語の数量表現コーパス

Koyano et al. [6] では，NPCMJ [7] から数量表現を含む文を抽出し，126 文に含まれる 278 件の数量表現にアノテーションを行うことで数量表現コーパスを構築した．この数量表現コーパスでは，数量表現を 2 つ以上含む文，否定表現または条件節を含む数量表現を 1 つ以上含む文に対して，助数辞，出現形式，用法のアノテーションを行なっている．

助数辞の分類　Koyano et al. [6] では，飯田 [1, 2] による分類辞，単位形成辞，計量辞と，奥津 [3] による順序数辞の 4 分類を採用している．各助数辞の例と数量表現コーパスに含まれる件数を付録 表 8 に示す．

分類辞には，助数辞単体では使用されないもの（「人」「頭」），複数の人やものに対して使用されるもの（「組」），汎用的な助数辞（「個」「つ」）が含ま

れる．単位形成辞は，何らかの容器を表す普通名詞（「箱」「セット」）であり，助数辞としてではなく単独で用いることが可能である．また，より大きなものの一部を表す助数辞（「切れ」）も含まれる．計量辞は，重さや距離を表す単位（「キロ」），長さを表す単位（「メートル」）など，測定のための単位を表す助数辞である．順序数辞は，時間（「日」「分」）や順序（「番」「位」）を表す助数辞である．順序数辞を含む数量表現は，修飾する名詞が文中に現れないことが多く，むしろ数量表現自体が名詞としての役割を果たすといった特徴がある．

数量表現の出現形式　数量表現の出現形式は，名詞を修飾する数量表現，動詞を修飾する数量表現によって文中に現れる位置が異なる．また，名詞を修飾する数量表現は，同じ状況を表す文でも，文脈によって名詞の前に位置する場合（「3 人の学生」）や後ろに位置する場合（「学生 3 人」）などがあり，出現形式は多様である．Koyano et al. [6] では，岩田 [5] が用いた 6 タイプと，動詞を修飾する数量表現，イベント名詞句を修飾する数量表現，修飾する名詞が文中に現れない数量表現，数量表現を修飾する数量表現，() 内の数量表現，イディオム的な数量表現について，6 タイプの出現形式を作成し，計 12 タイプに分類している．

数量表現の用法　数量表現の用法は，名詞を修飾する数量表現について，岩田 [5] が研究対象としていた 1 タイプに 3 タイプを追加した 4 タイプを作成し，動詞を修飾する数量表現については 4 タイプを作成している．さらにイベント名詞句を修飾する数量表現，数量表現を修飾する数量表現，イディオム的な用法についてのタイプを作成し，計 11 タイプに分類している．

3　実験

3.1　テストセット

Koyano et al. [6] の数量表現コーパスの 151 件の数量表現の助数辞をマスクし，テストセットを作成する．(4) は数量表現コーパス内に含まれる文の例であり，(5) は構築したテスト文の例である．数量表現コーパスは，(4) のように<num>タグで囲まれている数量表現に対して，助数辞，出現形式，用法がアノテーションされている．(4) の助数辞は「人」であるため，この部分を (5) のように [MASK] に置き換え，<num>タグが無い形式の文を作成する．

(4)　施設の周辺に滞在する外国人研究者や職員は常時約<num>３０００人</num>，その家族も含めると約１万人規模の人口増加となる．

(5)　施設の周辺に滞在する外国人研究者や職員は常時約３０００ [MASK]，その家族も含めると約１万人規模の人口増加となる．

さらに，各テスト文に対して，正解の助数辞と正解とは異なる助数辞（誤りの助数辞）のペアを作成する．(5) の正解の助数辞は，(4) の助数辞である「人」とし，誤りの助数辞は，同じ分類であるが (5) の [MASK] に入れると意味的に不自然となる助数辞とする．この例では，「本」を誤りの助数辞とする．誤りの助数辞は，[MASK] に入れると意味的に不自然になる助数辞を全て人手で選び，テストセットを構築する．

3.2　実験設定

3.1 項で構築したテストセットを用いて，汎用言語モデルが助数辞を正しく予測することができるかを分析する．マスクされた箇所に対する正解の助数辞と誤りの助数辞のそれぞれの予測確率を計算する．そして，誤りの助数辞よりも正解の助数辞に対して高い確率で予測した数の割合を正答率として算出して分析する．マスクされた箇所に入る助数辞を正しく予測しているかという実験は，自然言語推論タスクとは異なりファインチューニングを行わないで実験を行うため，より直接的に事前学習で何を学習しているのか評価できる手法である．

現在の標準的な汎用言語モデルが助数辞を正しく予測することができるかを評価するために，東北大 BERT[1] と早大 RoBERTa[2] の評価実験を実施した．

3.3　実験結果と考察

東北大 BERT と早大 RoBERTa を用いた評価実験の助数辞ごとの正答率を表 9，数量表現の出現ごとの正答率を表 4，用法ごとの正答率を表 5 に示す．(テストセットに含まれる各タイプの統計情報は付録 表 8 を参照)

助数辞ごとの正答率では，東北大 BERT と早大 RoBERTa のどちらも単位形成辞の正答率がやや低いものの，全体としては東北大 BERT は 89.00%，早大 RoBERTa は 94.04%と高い正答率だった．数量表

1）https://huggingface/cl-tohoku/bert-base-japanese-whole-word-masking
2）https://huggingface.co/nlp-waseda/roberta-base-japanese-with-auto-jumanpp

表 1 東北大 BERT が正しい助数辞よりも誤りの助数時を高い確率で予測した例．() 内はモデルの予測確率を示す．

	テスト文	正しい助数辞	誤りの助数辞
(a)	これは何を意味するかというと，ある一 [MASK] の産業分野は 必ず一社独占化の道を進む，つまり学問に基づいた本物の技術を作った企業 （多くの場合，得意技術の異なる複数の連合体）が勝つということである．	つ (13.72)	社（14.82）
(b)	帝国データバンク仙台支店が昨年１２月下旬に公表した景気見通し調査では， １４ [MASK] を「回復局面」とする企業は１８．４%．	年（12.49）	%（12.92）
(c)	チリは２ [MASK] 連続の再選が禁止され，４年ごとに大統領が代わる． チリは２期連続の再選が禁止され，４ [MASK] ごとに大統領が代わる．	期（17.82） 年（15.60）	年（19.28） 期（17.17）
(d)	国道３８号では赤信号で１ [MASK] ほど足止めを余儀なくされた．	分（13.28）	ヶ月（17.17）

表 2 早大 RoBERTa が正しい助数辞よりも誤りの助数時を高い確率で予測した例

	テスト文	正しい助数辞	誤りの助数辞
(e)	チリは２ [MASK] 連続の再選が禁止され，４年ごとに大統領が代わる．	期（19.74）	年（20.12）
(f)	一般的に言うと累積雨量が２００ [MASK] 以上降ると危険度は上がる．	ミリ（-3.56）	メートル（0.16）

現の出現形式，用法ごとの正答率は，東北大 BERT はイディオムのみ正答率が低く，その他のタイプでは 8 割以上の正答率だった．早大 RoBERTa の出現形式，用法ごとの正答率は，どのタイプも高かった．

東北大 BERT が正しい助数辞よりも誤りの助数辞を高い確率で予測した例を表 1 に示す．(a) は，数量表現の出現形式，用法がイディオムの例である．この文の [MASK] の箇所は「ある一つ」というイディオム的な数量表現であるため，「つ」が正解の助数辞である．この文の「一つ」という数量表現は，数詞を変更すると非文（「ある二つ」は非文）となるため，イディオムというタイプが付与されている．(a) では，マスクされている箇所の他に「一社」という同じ数詞「一」を持つ数量表現が含まれているため，マスクされている箇所の助数辞として「社」を高い確率で予測した可能性がある．(b) も同様に，誤りの助数辞「%」が文内のマスクされた箇所より後に現れる数量表現「18.4%」に含まれており，その助数辞を正しい助数辞よりも高い確率で予測していることから，テスト文に含まれる別の数量表現から助数辞を予測している可能性がある．

(c) の 1 つ目の例は，[MASK] に誤りの助数辞「年」が入った場合「2 年連続の再選が禁止され」となり，この節においては自然な文に見えるが，後続の節を含めた文では，意味が通らない文になってしまう（(c) の 2 つ目の例も同様）．このことから，東北大 BERT は長い文の意味（節の間の意味関係）を正しく捉えられていない可能性がある．

(d) は，正解の助数辞，誤りの助数辞ともに時間（期間）に関する助数辞であるが，人間なら「赤信号で 1 ヶ月も足止めされることはない」と容易に推測することができ，「赤信号で 1[MASK] ほど足止めを余儀なくされた」の助数辞は「分」もしくは長時間でも「時間」であると推測することができる．しかし，東北大 BERT は「ヶ月」を高い確率で予測していることから，文脈から時間に関する数量表現だということは予測できても，時間的な常識である時間の尺度まで考慮して推測することは難しいと考えられる．

表 3 助数辞ごとの評価実験の結果（正答率）

タイプ	東北大 BERT	早大 RoBERTa
ALL	89.40%	94.04%
分類辞	82.50%	92.50%
単位形成辞	75.00%	75.00%
計量辞	91.49%	93.62%
順序数辞	94.64%	98.21%

早大 RoBERTa が正しい助数辞よりも誤りの助数辞を高い確率で予測した例を表 2 に示す．(e) は表 1(c) の１つ目の例と同じテスト文である．早大 RoBERTa でも東北大 BERT と同じように，誤りの助数辞が文内のマスクされた箇所と異なる数量表現に含まれているとき，その助数辞を正しい助数辞よりも高い確率で予測することがあった．

(f) は，正解の助数辞より誤りの助数辞のほうが高い確率で予想されているが，どちらも予測確率が低かった．このテスト文では「です」「ます」といった助動詞や「が」「の」などの助詞の確率が高かった．このテスト文にはマスクされている箇所以外に数量表現がなく，別の数量表現から予測することができないため，助数辞が予測結果の上位にならない可能性がある．

表 4 出現形式ごとの評価実験の結果（正答率）

タイプ	東北大 BERT	早大 RoBERTa
Q ノ NC 型	95.83%	95.83%
N ノ QC 型	100.00%	100.00%
NCQ 型	89.47%	94.74%
NQC 型	100.00%	100.00%
デ格型	100.00%	85.71%
述部型	100.00%	93.33
QV 型	88.64%	97.73
NvCQ 型	100.00%	100.00%
N の脱落	77.78%	77.78%
QtQ 型	85.71%	85.71%
イディオム	50.00%	90.00%
(Q)	100.00%	100.00%

表 5 用法ごとの評価実験の結果（正答率）

タイプ	東北大 BERT	早大 RoBERTa
Q が N のカテゴリー情報を表すもの	91.67%	91.67%
Q が N を構成する要素の全体数を表すもの	100.00%	100.00%
Q が N を構成する要素の一部を表すもの	100.00%	100.00%
Q が N の属性や特徴を表すもの	94.87%	97.44%
V が行われた回数を表す Q	0.00%	100.00%
V が行われた期間を表す Q	92.86%	100.00%
V が行われた時間を表す Q	92.31%	96.15%
Nv を修飾する Q	100.00%	75.00%
イディオム	50.00%	90.00%
V の特徴を表す Q	90.00%	90.00%
Q を修飾する Q	85.71%	85.71%

4 関連研究

英語における数量表現の研究として，言語モデルの数量表現の扱いに関する調査 [8] がある．Cui et al. [8] は，多言語で事前学習された言語モデルが，英語における様々な数量表現を含む一般化量化子の振る舞いをどの程度捉えることができるかについて，一般化量化子の理解に特化したベンチマーク GQNLI を構築し，調査を行った．GQNLI で言語モデルを評価した結果，言語モデルの最高精度は 48%であった．NLI モデルや質問応答モデルの性能改善において，一般化量化子を捉えられないことが課題の一つとなっていることを示した．日本語における数量表現を扱った研究として，Narisawa et al. [9] による数量表現を含む自然言語推論を解くための実装と評価がある．Narisawa et al. [9] は，日本語の含意関係認識において数量表現が問題になる事例に焦点を当て分析を行い，数量表現の規格化のための実装と評価を行った．Narisawa et al. は，数量表現が出現する文ペアを 7 つのカテゴリに分類し，正しく含意関係を判定するために必要な処理について述べている．

マスク穴埋めタスクで汎用言語モデルの分析を行った先行研究として，英語では，事前学習済み言語モデルが否定表現と偽のプライミング（ひっかけ）の扱いに課題があることを示した研究 [10] がある．Kassner et al. [10] は，事前学習済み言語モデルが否定表現を理解しているかを分析するために，*Birds can [MASK]*，*Birds cannot [MASK]* といった，否定表現を含まない文と含む文を用いて，マスクされた箇所の上位 3 件の予測単語を調査した．その結果，どちらの文でも同じ単語（上記の例では *fly*）が上位となることから，言語モデルは否定表現を理解できていないことを示唆した．また，事前学習済み言語モデルがひっかけの影響を受けるかを分析するために，*Lexus is owned by [MASK]*，*Microsoft? Lexus is owned by [MASK]* といった，ひっかけを含めた文を作成し，マスクされた箇所の上位 3 件の予測単語を調査した．その結果，ひっかけを含む文では，ひっかけ自体やひっかけに関連する単語（*Microsoft*，*Google*）が正解の単語（*Toyota*）よりも高い確率で予測されている傾向があることから，言語モデルの単語の予測結果は，ひっかけの影響を受けていることを示した．

5 おわりに

本研究では，汎用言語モデルにおいて数量表現を含む推論の扱いに課題がある原因として，事前学習時に数量表現の振る舞いを正しく学習できていないという仮説を立てた．この仮説を検証するために，Koyano et al. [6] の数量表現コーパスを用いて，助数辞をマスクした文のテストセットを構築し，汎用言語モデルの評価実験を行なった．その結果，汎用言語モデルは 8 割以上の助数辞を正しく予測した．一方で，節の間の意味関係を考慮した単語の予測には，まだ課題があることを確認した．また，言語モデルの単語の予測結果は，文中に現れる他の数量表現や単語の影響を受けており，英語での研究において指摘のあった，事前学習済み言語モデルがひっかけの影響を受けるという現象が，日本語数量表現においても存在することを確認した．

今後，汎用言語モデルにおける数量表現の扱いの課題についてさらに調査し，言語学の理論に基づく数量表現の分類と推論の関係性や，推論における数量表現のひっかけの影響について分析を進める．

謝辞

本研究は JST CREST JPMJCR20D2，JST さきがけ JPMJPR21C8 の助成を受けたものである．

参考文献

[1] 飯田隆. 日本語と論理. NHK 出版, 2019.

[2] Takashi Iida. Japanese semantics and the mass/count distinction. **Chungmin Lee, Young-Wha Kim and Byeong-Uk Yi (eds.), Numerals Classifiers and Classifier Languages (Routledge)**, pp. 72–97, 2 2021.

[3] 奥津敬一郎. 拾遺日本文法論. ひつじ書房, 1996.

[4] 矢澤真人. 数量の表現. 金田一春彦, 林大, 柴田武（編）, 日本語百科事典. 大修館書店, 1988.

[5] 岩田一成. 日本語数量詞の諸相. くろしお出版, 2013.

[6] Kana Koyano, Hitomi Yanaka, Koji Mineshima, and Daisuke Bekki. Annotating Japanese numeral expressions for a logical and pragmatic inference dataset. In **Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation**, 2022.

[7] NINJAL. NINJAL Parsed Corpus of Modern Japanese. (Version 1.0). 2016. https://npcmj.ninjal.ac.jp/.

[8] Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In **Proc. of NAACL**, 2022.

[9] Katsuma Narisawa, Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. Is a 204 cm man tall or small? acquisition of numerical common sense from the web. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 382–391, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7811–7818, Online, July 2020. Association for Computational Linguistics.

表 6　早大 RoBERTa の予測結果の上位 10 件に正解の助数辞が含まれない例

	テスト文	正しい助数辞	上位の助数辞（RANK）
(k)	ちなみに金板の百人一首は１ [MASK] で８０万円である．	セット	首 (1) 人 (9)
(l)	そんな仕事は三 [MASK] とかかりません．	日	の (6) です（7）

表 7　東北大 BERT の予測結果の上位 10 件に正解の助数辞が含まれない例

	テスト文	正しい助数辞	上位の助数辞（RANK）
(i)	研究すべき材料は三種類で，それを五つのちがった温度で，各十回測る， という風になっておれば，決して四 [MASK] はやってみない．	種類	回 (1) 度 (2)
(j)	そこには，ユダヤ人のきよめのならわしに従って，それぞれ 四，五 [MASK] もはいる石の水がめが，六つ置いてあった．	斗	つ (1) 人 (3)

表 8　各助数辞の例と数量表現コーパスに含まれる件数

タイプ	例	件数
分類辞	人，頭，組，個	62
単位形成辞	瓶，箱，袋，セット，切れ	11
計量辞	キロ，メートル	106
序数辞	日，分，番，位	108

A　付録：上位 N 件の予測結果

汎用言語モデルの予測結果の上位 1 件，上位 3 件，上位 5 件，上位 10 件に正解の助数辞が含まれる確率を表 10 に示す．どの結果も早大 RoBERTa が東北大 BERT よりも正答率が高かった．東北大 BERT の予測結果では，上位 10 件で予測結果に正解の助数辞が含まれる確率が 90%を超えるのに対し，早大 RoBERTa の予測結果では，上位 3 件で予測結果に正解の助数辞が含まれる確率が 90%を超えるという結果となった．

表 9　助数辞，出現形式，用法ごとの統計情報

タイプ	件数
ALL	151
分類辞	40
単位形成辞	8
計量辞	47
順序数辞	56
Q ノ NC 型	24
N ノ QC 型	4
NCQ 型	19
NQC 型	5
デ格型	7
述部型	15
QV 型	44
NvCQ 型	1
N の脱落	9
QtQ 型	7
イディオム	10
(Q)	6
Q が N のカテゴリー情報を表すもの	36
Q が N を構成する要素の全体数を表すもの	1
Q が N を構成する要素の一部を表すもの	3
Q が N の属性や特徴を表すもの	39
V が行われた回数を表す Q	1
V が行われた期間を表す Q	14
V が行われた時間を表す Q	26
Nv を修飾する Q	4
イディオム	10
V の特徴を表す Q	10
Q を修飾する Q	7

表 10　上位 N 件に正解の助数辞が含まれる確率

RANK	東北大 BERT	早大 RoBERTa
上位 1 件	70.86%	78.15%
上位 3 件	84.11%	90.07%
上位 5 件	87.42%	92.05%
上位 10 件	90.07%	92.72%