言語モデルを用いた 日本語の語順評価と基本語順の分析

栗林 樹生 [1,3*] 伊藤 拓海 [1,3*] 内山 香 [1] 鈴木 潤 [1,2] 乾 健太郎 [1,2]
[1] 東北大学 [2] 理化学研究所 AIP センター [3]Langsmith 株式会社
{kuribayashi, t-ito, jun.suzuki, inui}@ecei.tohoku.ac.jp
kaori.uchiyama.r1@dc.tohoku.ac.jp

1 はじめに

本研究では，日本語の語順の適切さに対する自動評価と評価傾向の分析を行う．理解しやすい日本語の文を書くためには語の配置に注意する必要があり，語順の適切さの自動評価はライティング支援などへ応用できると考えられる．例えば，基本語順から逸脱した文においては読み時間や容認性判断時間に遅れが出ること [3, 5–7] や, 可読性の指標として主語，述語，目的語間の距離などを考慮することの重要性が示唆されている [9].

語順の妥当性を評価する手段として，本論文では言語モデルを用いる．言語モデルは語順をモデリングしているため，語順の適切さの評価に用いることは妥当であると考えられる．図 1のように各用例に対してかき混ぜ文を作成し，言語モデルスコアでランキングすることで適切な語順を求めた．対象の文に対して，文の語順を入れ替えをたものをかき混ぜ文と呼ぶ．

始めに語順判定タスクを作成し，言語モデルが適切な語順をどの程度の精度で識別可能か検証した．本実験において言語モデルは人間と同程度の精度で正しい語順を識別可能であることが分かった．この結果から，言語モデルが日本語の自然な語順を正しく識別できるという仮定をおけることが示唆された．

次に，言語学の分野で示唆されている語順に関する諸仮説に対して言語モデルがどのような傾向を示すか網羅的に分析した．本研究では，特にこれまで研究が成されてきた，二重目的語構文における目的語の語順と副詞の位置に焦点を当てる [7, 8, 10]．これまでに成されていない数理的なモデルを用いた分析を行うことで，それらの仮説を支持する (支持しない) 新たな根拠を提示することができる．言語モデルを用いた分析では前処理やフィルタリング (複雑でない文のみを用いるなど) をしていない大量の生テキストから学習した傾向を用いるため，より一般的な日本語の傾向を踏まえることができることや，笹野ら [10] のように自動的に用例を収集する際の述語項解析器のミスの影響を受けにくいことなどの利点も考えられる．また，言語モデルがどの程度語順に対する諸言語現象を捉えているかという，言語モデル自体の性質の理解にも繋がると考えられる．

分析の結果から，語順に関する仮説に対して言語モデルが既存研究と非常に似た傾向を示すことを確認した．また，今まで大規模な言語資源を用いて調査できていなかったヲ格が非明示的に出現した場合の傾向も分析した．

本研究の貢献は以下の通りである．

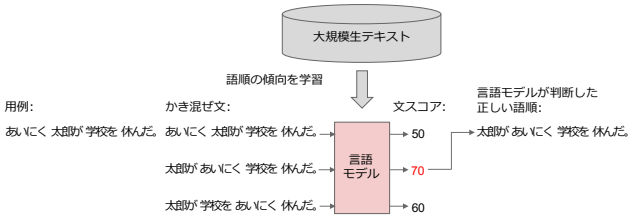

図1: 言語モデルによる適切な語順の判定.

- 言語モデルが適切な語順をどれほど識別できるかを検証し，人間と同程度に正しい語順を判別できることが分かった．
- 日本語の基本語順の分析に言語モデルを用いるという方法論を初めて提案した．
- 日本語二重目的語構文や副詞の語順における多くの仮説に対して，本分析結果が既存研究と似た傾向を示すことが分かったが，一部の仮説では既存研究と異なる傾向も示した．
- 大規模な言語資源を用いた非明示的なヲ格に対する分析を初めて行った．
- 副詞の分析で用いられた用例とかき混ぜ文において，文に対する人間の容認速度と言語モデルが算出した文の生成確率に 0.670 の相関がみられた．

2 語順分類実験

本実験を通して，言語モデルの正しい語順を識別する能力に関して調査する．京都大学テキストコーパス Version 4.0 に出現する文を用い，コーパスに出現した文とかき混ぜ文を識別させるタスクを定義した．本コーパスは毎日新聞の記事から抽出したものであり，コーパスに出現する文は適切な語順で書かれていると仮定し，かき混ぜ文は基本語順から逸脱した不自然な文と仮定した．かき混ぜ文を作成する際に適用したかき混ぜ操作はセクション2.1に示す．

複雑な文を除外するため，以下の条件を満たす文を用いた．

- 文節数が 5 以下である．
- 文内に動詞が 2 つ以上存在しない．
- 係り受け関係に注目した際に兄弟関係となるチャンクが存在し，さらにそれらが格助詞か副詞的名詞を伴っている．
- 括弧などの特定の記号が出現する文は除く．

したがって，節を越える移動 (長距離かき混ぜ操作) などは本実験の対象外とする．

2.1 かき混ぜ操作

かき混ぜ文を作る処理について説明する．本研究では，以下の現象が起きた文をかき混ぜ文とみなす．[ ] 内のチャンクが

*本研究の主要な貢献は第一著者と第二著者によるものであり，両者の貢献は同等である．

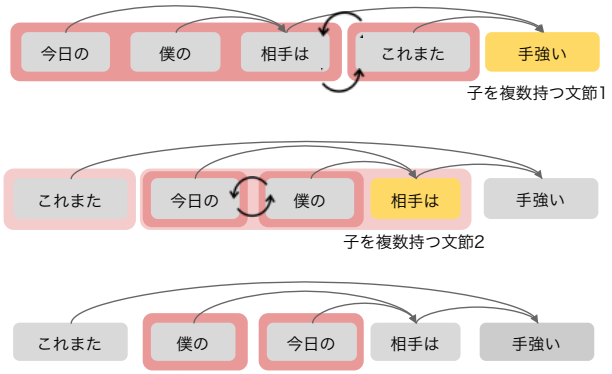

図2: かき混ぜ操作による基本語順から逸脱した文の作成方法.

(gap) の位置から移動した.

- 格助詞を伴う文節の出現位置の変更
  (1) [愛情を] 言葉に (gap) 感じる.
- 副詞の出現位置の変更
  (2) (gap) 順子が [不幸にも] 右足を骨折した.

かき混ぜ操作の概要を図 2に示す. コーパスに付与されている形態素・構文情報を用い, 係り受け関係の子を複数持つ文節が存在した場合, 子同士の位置をランダムに入れ替えた. 図 2内の一番上の操作のように, 入れ替え操作の対象となる子 (“相手は”) が子孫 (“今日の” と “僕の”) をもつ場合, それらを一つのまとまり (“今日の僕の相手”) とみなして動かした[*1].

2.2 実験設定

コーパス中に実在する文とそのかき混ぜ文のペアに対して, どちらが実在する語順の文かを当てる 2 値分類問題を解いた. ある文に対するかき混ぜ文は複数作成可能な場合があるが, その場合は作成したかき混ぜ文の中から 1 つをランダムに採用した. 実在する文とそのかき混ぜ文のペアを 500 ペア使用した.

CNN ベースの言語モデル[*2]を用い, 未知語を避けるためキャラクターベースで学習した. Web30 億ページから抽出した 90 億文のうち, ランダムに選ばれた 1.6 億文を言語モデルの学習に用いた. 与えられた 2 文に対して言語モデルを用いて文の生成確率を求め, 生成確率が高い文を言語モデルが支持する語順とみなした.

文スコアの求め方に関して以下の 3 つの方法を比較した.

1. 順方向言語モデル: 順方向言語モデルによる生成確率
2. 逆方向言語モデル: 逆方向言語モデルによる生成確率
3. 双方向言語モデル: 順方向言語モデルが求めた生成確率と逆方向言語モデルが用いた生成確率の平均

2.3 結果

表 1に結果を示す. 双方向の言語モデルを用いたスコアリングが最も良い性能を示した. 人間のスコアは本論文の共著者を含む 3 名が本タスクを解いた結果の平均である. 人間と同程度の識別精度を示し, 言語モデルは語順の適切さを正しく判断するという仮定をおくことの確からしさが示唆された. 以降の分析では, 言語モデルスコアとして双方向言語モデルスコアを用いる.

[*1] 入れ替え操作の対象となる子が子孫を持ち, 子孫から出る係り受け関係が他の係り受け関係と交差している場合は, 対象とした子の直前に子孫らを位置させた. また, 前向きの係り受け関係がある文は対象外とした.

[*2] https://github.com/pytorch/fairseq

表1: 各アプローチの予測精度.

Model	正答率
順方向言語モデル	83.4
逆方向言語モデル	81.6
双方向言語モデル	84.4
人間	83.1

3 日本語語順に関する仮説の網羅的分析

既存研究で示唆されている諸仮説に対して, 言語モデルがどのような傾向を示すか分析した. 既存研究による分析方法とは異なる数理的なモデルによる分析においても同じ検証結果が得られた場合, それらの仮説を支持する (支持しない) 新たな根拠を提示できたこととなる. また, 笹野ら [10] の分析ではニ格とヲ格の両方が明示的に出現し構文的曖昧性のない簡潔な文のみを用例として採用しているため, 数えているデータの性質にバイアスがかかっている可能性があるが, 言語モデルによる分析では Web 上に書かれている大量の文をそのまま利用して語順の傾向を学習しているため, 多くの人によって実際に書かれている日本語に近い傾向を反映していると考えられる.

以下の仮説を検証する.

1. 動詞によらず基本語順は「にを」である [1].
2. 省略されにくい格は動詞の近くに位置する傾向がある [10].
3. 動詞のタイプが基本語順に影響を与える. [4]
4. 二格名詞の有生性が基本語順に影響を与える [2, 4, 10].
5. 対象の動詞と高頻度に共起するヲ格名詞, ニ格名詞は動詞の近くに出現しやすい [7, 10].
6. ヲ格が副助詞「は」や副助詞「も」を伴って出現した場合, もともとの「をに」率よりも,「はに」率や「もに」率が上がる.
7. 副詞の基本語順は [8] の示したとおりである.

言語モデルによる分析結果と先行研究を厳密に比較するため, 先行研究時に用いられている用例と同様のものを用いた. 図 1に示すように, 既存研究で用いられている各用例に対して言語モデルに適切な語順を予測させ, それらの結果を言語モデルが予測した結果とした.

3.1 動詞ごとの「をに」率

図 3に分析結果を示す. 各プロットが各動詞の分析結果に対応しており, 横軸が笹野ら [10] が求めた「をに」率, 縦軸が各動詞の用例に対して言語モデルが良いとみなした語順における全体的な「をに」率である. 各動詞ごとの言語モデルが良いと判断した「をに」率と笹野ら [10] が求めた「をに」率は相関係数 0.89 で強く相関していた. また,「をに」率が非常に高い動詞が存在することから, 動詞によらず基本語順は「にを」であるという Hoji ら [1] の仮説は妥当であると考えにくい.

3.2 省略されにくい格は動詞の近くに位置する

図 3に, 動詞ごとの二格だけ出現した用例の割合と各動詞の「をに」率の関係を示す. 笹野ら [10] は, 省略されにくい格は動詞の近くに位置する傾向があることを示唆していたが, 言語モデルもこの傾向を示した. 動詞ごとの二格だけ出現した用例の割合と各動詞の「をに」率の間の相関係数は, 笹野ら [10] が 0.391 であったのに対して, 言語モデルを用いた分析では 0.399

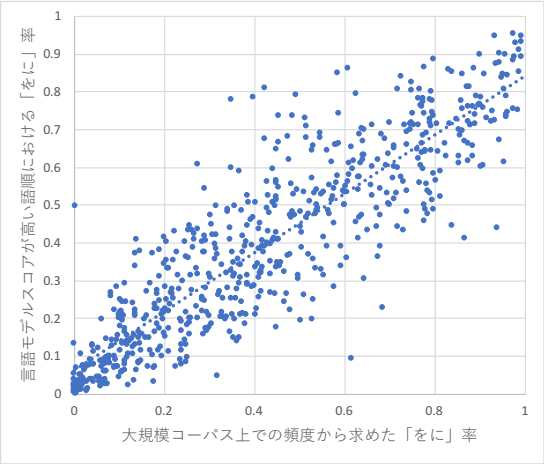

図3: 動詞ごとの「を/に」率．図中の直線は線形回帰直線を表す．

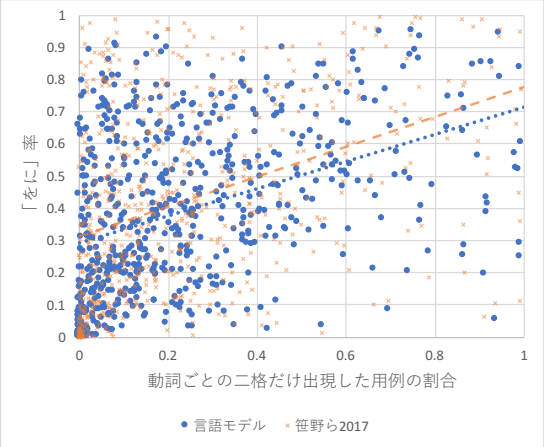

図4: 動詞ごとの二格だけ出現した用例の割合と各動詞の「を/に」率．図中の直線は線形回帰直線を表す．

であり，概ね同様の結果となった．

3.3 動詞のタイプと基本語順

表 2に show タイプと pass タイプの各動詞に対して求めた「を/に」率とタイプごとの Macro 平均を示す.[*3]．ウィルコクソンの順位和検定で 2 群の「を/に」率の分布を比較し，Show タイプと Pass タイプで優位な差は見られなかった (p=0.354)．この結果は笹野ら [10] と一致する．

3.4 二格名詞の有生性と「を/に」率

ヲ格名詞のカテゴリが『人工物-その他』である用例に限定した上で，笹野ら [10] と同様に 126 動詞を分析対象とし，二格名詞のカテゴリが「人」である場合と「場所-施設」である場合の「を/に」率の違いを調べた．笹野ら [10] は，二格名詞のカテゴリが「人」である場合と「場所-施設」である場合で「を/に」率に優位な違いが出ることを示していたが，言語モデルによる分析では両者にはほとんど差がでなかった (表 3)．この結果からは，言語モデルが格要素の有生性を考慮できていない，笹野ら [10] が指摘しているように収集した用例の中にノイズとなる文が含まれている，二格カテゴリの有生性が二重目的語構文の語順に影響を及ぼさないなどの可能性が示唆される．

[*3]笹野ら [10] と異なり，文献 [10] が分析対象に追加した「言付ける」，「知らせる」，「返す」は分析対象に含めなかった．

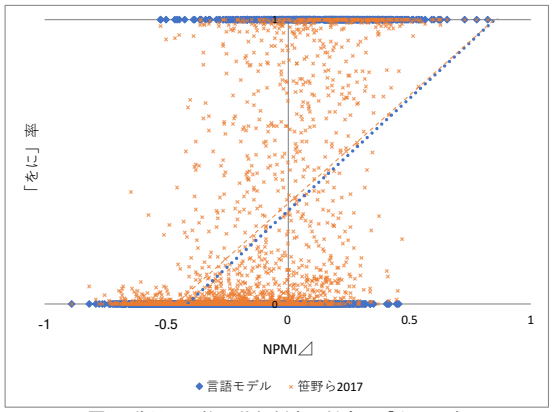

図5: 動詞と二格の共起頻度に対する「を/に」率

3.5 動詞と共起しやすい語は動詞の近くに出現する

動詞と二格名詞の共起頻度と「を/に」率について言語モデルを用いた分析結果を図 5に示す．各プロットは各用例の「を/に」率に対応し，青のプロット (丸) は言語モデルの予測結果 (言語モデルが「を/に」語順を支持した場合縦軸の値は 1,「に/を」語順を支持した場合縦軸の値は 0 とした.) ，オレンジのプロット (バツ印) は笹野ら [10] の分析結果に対応する．横軸 ΔNPMI は，笹野ら [10] と同様の尺度を用いており, 対象の動詞に対してヲ格に出現している名詞よりも二格に出現している名詞のほうが共起しやすい度合いを表している．笹野ら [10] が報告した，ニ格名詞と動詞，ヲ格名詞と動詞の NPMI の値の差と「を/に」語順の割合の相関係数は全体で 0.567，我々の言語モデルを用いた分析では 0.539 であり，共に対象の動詞と高頻度に共起するヲ格名詞，ニ格名詞は動詞の近くに出現しやすいことを支持する結果となった．

3.6 ヲ格が副助詞「は」や副助詞「も」を伴って出現した場合「を/に」語順をとる

ヲ格が副助詞「は」や「も」を伴って出現した場合，それらは文の主題となる可能性が高く，主語として文の先頭寄り (動詞から離れた位置) に登場するのではないかという仮説を立てた．このような非明示的な格に対する大規模な言語資源を用いた分析は行われていない．例えば大規模なコーパスに対して解析をかけた後収集した用例を数えるという笹野ら [10] のアプローチは，非明示的な格に対する述語項解析精度が低いことから大規模で高品質な用例の収集が困難であり，本仮説の検証には向いていないと考えられる．

全ての格助詞「を」は副助詞「も」や「は」に置き換えることができるという仮定をおき，笹野ら [10] の用例に対して格助詞「を」を「は」に置き換えた用例群 (「はに」コーパス) と「も」に置き換えた用例群 (「もに」コーパス) を作成した．「はに」コーパスと「もに」コーパスに対して 3.1章と同様の分析をした結果を図 6に示す．ヲ格が明示的に出現した時 (青点) に比べて，ヲ格が副助詞「も」を伴って現れたときのほうが，更に副助詞「は」を伴って出現したときのほうが「を/に」率が高くなる傾向にあった．特に,「にを」率が極端に高くない動詞においてこの傾向は顕著であった．また，笹野ら [10] が報告した各動詞の「を/に」率と,「もに」コーパスにおける言語モデルが予測した各動詞の「を/に」率との相関は 0.851,「はに」コーパス

表2: 動詞と二格カテゴリに対する「をに」率.

Show タイプ			Pass タイプ					
動詞	言語モデル	笹野ら (2017)	動詞	言語モデル	笹野ら (2017)	動詞	言語モデル	笹野ら (2017)
知らせる	0.897	0.958	戻す	0.638	0.771	落とす	0.375	0.351
預ける	0.466	0.399	泊める	0.717	0.748	漏らす	0.254	0.332
見せる	0.317	0.301	包む	0.373	0.603	浮かべる	0.361	0.255
被せる	0.189	0.256	伝える	0.477	0.522	向ける	0.182	0.251
教える	0.251	0.235	乗せる	0.541	0.496	残す	0.079	0.238
授ける	0.273	0.186	届ける	0.474	0.491	埋める	0.429	0.223
浴びせる	0.069	0.177	並べる	0.473	0.481	混ぜる	0.415	0.200
貸す	0.138	0.118	ぶつける	0.268	0.436	当てる	0.225	0.185
着せる	0.078	0.113	付ける	0.148	0.368	掛ける	0.135	0.108
-	-	-	渡す	0.277	0.362	重ねる	0.128	0.084
Macro 平均	0.298	0.305	Macro 平均				0.334	0.361

表3: 動詞と二格カテゴリに対する「をに」率.

二格カテゴリ	言語モデル	笹野ら (2017)
人	0.431	0.384
施設	0.447	0.468

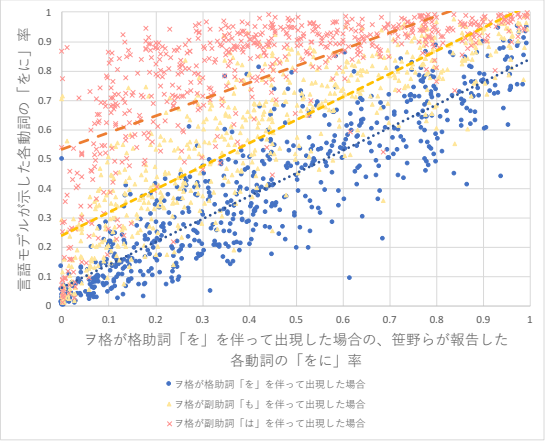

図6: ヲ格が非自明に出現した際の「をに」率．図中の直線は線形回帰直線を表す.

表4: 既存研究における基本語順と，言語モデルの示した自然な語順の傾向.A は副詞，S は主語，O は目的語，V は動詞を示す.

副詞類	言語モデルによる自然さ	人間の容認速度による自然さ [8]
陳述	ASOV>SAOV>SOAV	ASOV>SAOV>SOAV
時	ASOV=SAOV>SOAV	ASOV=SAOV>SOAV
様態	**SAOV>SOAV**>ASOV	SAOV=SOAV>ASOV
結果	SOAV=SAOV>ASOV	SAOV=SOAV>ASOV

においては 0.700 であり，ヲ格が明示的に出現する際の各動詞の「をに」率と，ヲ格が非明示的に出現した際の「をに」率は相関することがわかった．ヲ格が副助詞を伴って現れても「をに」率が低いままの動詞として，「差し伸べる」，「隠せる」，「表する」などが挙げられる.

3.7 副詞類の語順の傾向

小泉ら [8] の分析で用いられた用例に対して，モデルに適切な語順を判断させた．表 4に小泉ら [8] の分析結果と言語モデルを用いた分析の結果を示す．言語モデルの結果の語順間の不等号に関しては，符号検定で順位の差に有意差が見られる ($p<0.05$) ことを基準とした．小泉ら [8] と概ね一致する結果となったが，様態の副詞に関して小泉ら [8] は SAOV と SOAV の両方が基本語順であるとみなしていたが，言語モデルでは SOAV よりも SAOV の方が自然であると優位に判断された．様態の副詞を含む文の例を表 5に示す．また，各用例に対する言語モデルスコアと小泉ら [8] が報告した人間の容認速度を比較したところ 0.670 で相関が見られた.

表5: 様態の副詞を用いた文のうち，言語モデルによって副詞を目的語の前に置いたほう (SAOV) が自然であると判断された文の例.

自然であるとされた SAOV 語順	SOAV 語順
健二が難なく相手をやっつけた。	健二が相手を難なくやっつけた。
順子がのろのろと車を運転した。	順子が車をのろのろと運転した。
健二がこっそり玄関を開けた。	健二が玄関をこっそり開けた。

4 おわりに

本研究では，日本語の語順の適切さに対する言語モデルを用いた自動評価と評価傾向の分析を行った．語順分類実験を通して言語モデルの識別能力を調べ，人間と同程度の精度で正しい語順を識別可能であることを確認した．また，基本語順の分析に言語モデルを用いるという新たな方法論を提示し，語順に関する諸仮説の網羅的な検証を行った．これまでの手法では困難であった大規模な言語資源を用いた非明示的なヲ格に対する分析も行うことができた．今回は文字レベルの言語モデルで日本語の基本語順について分析を行ったが．形態素レベルの言語モデルやマスクを用いた言語モデルなどの異なるタイプの言語モデルや，語順に限らない言語現象に関しても分析を検討したい.

謝辞

本研究は東北大学 Step-Qi スクールの支援を受けた.

参考文献

[1] Hajime Hoji. “Logical form constraints and configurational structures in Japanese.” In: (1986).
[2] Atsushi Ito. “Argument structure of Japanese ditransitives”. In: *Nanzan Linguistics Special* 3 (2007), pp. 127–150.
[3] Masatoshi Koizumi and Katsuo Tamaoka. “Cognitive processing of Japanese sentences with ditransitive verbs”. In: *Gengo Kenkyu (Journal of the Linguistic Society of Japan)* 2004.125 (2004), pp. 173–190.
[4] Mikinari Matsuoka. “Two Types of Ditransitive Consturctions in Japanese”. In: *Journal of East Asian Linguistics* 12.2 (2003), pp. 171–203.
[5] Reiko Mazuka. “Costs of scrambling in Japanese sentence processing”. In: *Sentence processing in East Asian languages* (2002), pp. 167–188.
[6] Katsuo Tamaoka et al. “Priority information used for the processing of Japanese sentences: Thematic roles, case particles or grammatical functions?” In: *Journal of Psycholinguistic Research* 34.3 (2005), pp. 281–332.
[7] 中本敬子, 李在鎬, and 黒田航. “日本語の語順選好は動詞に還元できない文レベルの意味と相関する”. In: 認知科学 13.3 (2006), pp. 334–352.
[8] 小泉政利 and 玉岡賀津雄. “文解析実験による日本語副詞類の基本語順の判定”. In: 認知科学 13.3 (2006), pp. 392–403.
[9] 祖国威, 吉村裕美子, and 加納敏行. “構文的特性に着目した可読性診断技術”. In: 東芝レビュー 66.4 (2011), pp. 51–55.
[10] 笹野遼平 and 奥村学. “大規模コーパスに基づく日本語二重目的語構文の基本語順の分析”. In: 自然言語処理 24.5 (2017), pp. 687–703.