主題化における人間と言語モデルの対照

藤原吏生 [1]　栗林樹生 [1,2]　徳久良子 [1]　乾健太郎 [1,3]
[1] 東北大学　[2]Langsmith 株式会社　[3] 理化学研究所
riki.fujihara.s4@dc.tohoku.ac.jp
{kuribayashi,tokuhisa,kentaro.inui}@tohoku.ac.jp

概要

人間は文脈から自明な内容を省略したり，先行する文脈に関連する情報を文頭に配置したりすることで，文章を滑らかに繋ぐ．本研究では，このような文章生成における談話レベルの選好について，ニューラル言語モデルが人間らしい振る舞いをしているかを調査する．特に日本語における項の主題化の選好に焦点を当て，クラウドソーシングを用いて大規模に人間の選好データを収集し，人間と言語モデルの選好を対照する．本実験の範囲では，主題化について言語モデルが人間とは異なる選好を示し，談話レベルの現象について言語モデルが人間と異なる汎化を行なっている可能性が示された．

1　はじめに

ニューラルモデルはデータのみから人間らしい言語の汎化を達成するのかという認知科学的な関心のもと，近年進展を遂げたニューラル言語モデルの有する言語知識が分析されてきた [1, 2, 3, 4, 5, 6]．文章の産出といった言語活動において文を超えた談話的な側面は無視できない一方，言語モデルの分析では統語的知識などが関心の的となり [7, 8]，談話的知識については文の並び替え能力など粗いレベルでの分析にとどまることが多い．

本研究では談話の一側面である主題構造 [9] に焦点を当て，言語モデルが有する談話レベルの選好について，人間と対照しながら統制的に調査する．主題は，「心理的主語：話者が伝えたいことの算出にとりかかる時にまず心の中にもつもの」と定義され [10]，ある要素を主題にするか（主題化するか）という判断は，センタリング理論 [11, 12] や情報の新旧 [13] などと関わりがあり，文脈を考慮する必要があることが知られている．

例えば以下の 2 文は「花瓶」が主題化されているかという観点で異なり，「花瓶を昨日割った。」という文の続きとしては，文 (1a) よりも文 (1b) を用いることが考えられる．

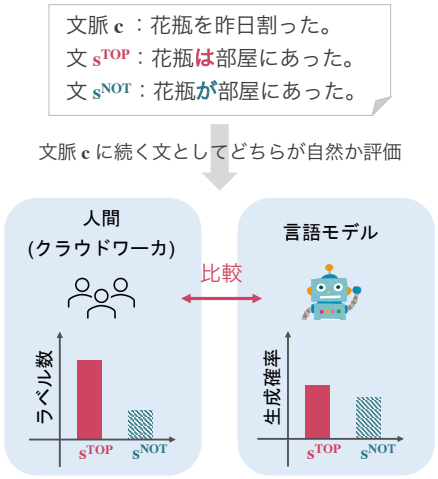

図 1　主題化判断における人間と言語モデルの比較．特定の文脈において，ある項が主題化された文 s^{TOP} と，主題化されていない文 s^{NOT} のどちらが続くのが自然かを判断させる

(1)　a.花瓶が部屋にあった。
There was a vase in the room.
b.花瓶は部屋にあった。
The vase was in the room.

このような文脈依存な判断について，言語モデルの選好を分析する．また，主題化の選択は日本語学習者にとって容易でなく [14]，主題化のモデリングは日本語の主題選択に関する自動判定や，文章執筆支援といった応用的出口にも繋がる．

具体的には，文の主題がとりたて助詞で明示される日本語を用い，統制的に文ペアを作成することで主題構造の観点を切り分けてモデルを分析する．人間の文脈依存な選好に関してクラウドソーシングを用いて大規模にデータを収集した後，主題化に関する選好を言語モデルと人間とで比較した（図 1）．人間の判断が文脈依存で変化する事例において，言語モデルは文脈非依存な振る舞いを示し，日本語の主

題化について人間とは異なる汎化をしていることが示唆された．本研究で作成したデータセットは公開する[1]．

2 データセット作成

本研究では，述語の項のうち，特に主格，与格，対格の主題化を対象とする．以下に，ある項が主題化された文とされていない文のペアを例示する．

主格の主題化の例

(2) a.花瓶が部屋にあった。
b.花瓶は部屋にあった。

与格の主題化の例

(3) a.飛行機で沖縄に向かった。
b.沖縄には飛行機で向かった。

対格の主題化の例

(4) a.沖縄で会議を開催した。
b.会議は沖縄で開催した。

実験では，特定の文脈のもと，続く文として項を主題化した/しない文のどちらが自然かを人間と言語モデルに問い，選好を比較する．

2.1 データ収集

NAIST Text Corpus（NTC） [15] に主題化に関する判断のアノテーションを行う．述語項構造アノテーションを利用し，以下の条件を全て満たす項（主格・与格・対格）を対象とした．[2]

- 文内で最も後方に出現する能動態の動詞の項である
- 項がとりたて助詞「は」を伴うか，各格に対応する格助詞（が・に・を）を伴って表出している
- 与格/対格を対象とする場合は，対応する述語が対格/与格の項を持たない
- 文中で対象の項以外が主題化されていない
- 項が各文章の先頭から 2 文目以降に出現する

上の条件を満たしたある項 a_i を含む文 s_i について，2 節で示したような主題構造の観点でのみ異なる文ペアを作成した．文 s_i で対象とする項 a_i が主題化されていた場合（とりたて助詞「は」を伴う場合）は，とりたて助詞を格助詞に変え，項を基本語順位置[3]に移動させることで，主題化が生じていない無標の文を作成した．逆に，項が主題化されていなかった場合は，とりたて助詞「は」を付与して文頭に移動することで，主題化が生じている有標の文を作成した．結果としてある項 a_i について，(文脈 c_i，項が主題化された文 s_i^{TOP}，項が主題化されていない文 s_i^{NOT}) の 3 つ組が収集された．データの例を表 1 に示す．

2.2 クラウドソーシング

各項 a_i が主題化される尤もらしさを定量化し，さらに主題化の判断における文脈 c_i の影響を測るために，(i) 文脈 c_i を見せない場合[4]と (ii) 文脈 c_i を見せる場合[5]の 2 つの設定で，文 s_i^{TOP} と文 s_i^{NOT} のどちらが自然かを判断させた．「どちらでも良い」という選択肢も提示し，どうしても判断がつかない場合のみ選択するよう指示した．また両設定では異なる作業者がアノテーションを行った．また 2.1 節で施した機械的な文ペアの生成により，非文が生成される可能性があるため，どちらの文も日本語の文として明らかな違和感がないことも作業者に確認し，非文だと判断された文が含まれる事例はデータから除外した．

アノテーションには Yahoo!クラウドソーシング[6]を用いた．事前に同様のタスクをトライアルとして複数回実施して優秀な作業者を選別し，さらにチェック設問を用いて不適切なラベルを付与する作業者を適宜除外した．この時点で，各格 × 各設定（文脈を見せる場合と見せない場合）に対して 8 人ずつのラベルを収集した．その後，MACE [16] を用いて各作業者の信頼度を計算し，信頼度下位 30%の作業者を除外したのち，3 人以下のラベルしか付与されていないデータポイントも除外した．最終的に得られたデータセットの統計量を表 2 に示す．最終的に，主格のアノテーションでは 164 名の，与格・対格のアノテーションでは 247 名の作業者の作業結果を採用した[7]．なお，主格のデータ収集についての詳細は，Fujihara ら [17] を参考にされたい．

1） https://github.com/rk-fujifuji/lm_topicalization
2） 主格の場合のみ，「対象とする格以外が主題化されていない」という条件は用いず，また事例を十分な量収集できたため，作業者の先行文脈を読む負荷を考慮して文章の先頭 2 から 4 文目までを対象とした．
3） 脱主題化時に，主格の場合は位置は変えず，与格・対格については述語の直前に移動した．なお，与格と対格が同時に出現する文は省いているため，基本語順において与格と対格のどちらが先かという議論は生じない．
4） 特定の文脈を想定しないよう作業者に指示した．
5） 作業者の負担を軽減するために，文脈 c が 4 文以上からなる場合は，後ろから 3 文目までを見せた
6） https://crowdsourcing.yahoo.co.jp/
7） トライアルも含め，チェック設問に正解した作業者に対して 1 タスク (約 10 分) あたり 160 円を報酬として支払った

表 1　データセットの例（機械的に作成した文の文頭には*を付与した）

文脈 c	文 s^{TOP}	文 s^{NOT}
外交研究家、清沢洌氏の日記によると、前夜、帝国ホテルで沢田外務次官と会食。	**沢田次官は**「『東洋経済』に外相と首相の仲が悪いように書いてある。大臣は迷惑している」と不満を述べた。	***沢田次官が**「『東洋経済』に外相と首相の仲が悪いように書いてある。大臣は迷惑している」と不満を述べた。
二月に公開が予定されている映画「フランケンシュタイン」の監督・主演をしたケネス・ブラナーが来日した。ブラナーは「から騒ぎ」などのシェークスピア原作映画を製作するなど、活発な創作活動を続けている。	**「フランケンシュタイン」は**フランシス・フォード・コッポラの製作、ロバート・デ・ニーロの主演という面でも見逃せない。	*フランシス・フォード・コッポラの製作、ロバート・デ・ニーロの主演という面でも**「フランケンシュタイン」を**見逃せない。
千畳敷付近は、星空の名所。「降るような星空でした」と星野さんが言うように、この夜も満天の星空が楽しめたという。星野さんらは三日午前中に木曽駒ケ岳に登頂。	***昼食は**、昼ごろに宝剣山荘に着き、全員で済ませた。	昼ごろに宝剣山荘に着き、全員で**昼食を**済ませた。
日米科学技術協力協定に基づく日米合同高級委員会が十二日、東京で開かれ、オゾン層破壊の解明につながる北極圏上空の大気観測を日米共同で行うことを新たに決めた。	**委員会には**日本から田中真紀子科学技術庁長官、米側からジョン・ギボンズ大統領補佐官らが出席した。	*日本から田中真紀子科学技術庁長官、米側からジョン・ギボンズ大統領補佐官らが**委員会に**出席した。

表 2　データセットの統計量

格	インスタンス数	ラベル数
主格 [17]	1,355	16,133
与格	256	2,861
対格	543	6,116

2.3　文脈依存セットの作成

2.2 節で収集したデータの中には，文脈の影響が見られない，すなわち文脈を見せた場合と見せない場合とで作業者の選好がほとんど変化しないインスタンスが確認された．また，「は」と「が」の使い分けは，談話レベルの主題構造以外の観点の作用が無視できないという調査もある [18, 19]．そこで，言語モデルの文脈（非）依存な振る舞いを精緻に分析するために，以下の条件のいずれかを満たすインスタンスを文脈依存セットとして抽出した．

- NTC 上と文脈を見せた場合のアノテーションで主題化された文が選ばれた，かつ文脈の有無による選好の変化の大きさが上位 25%に属する（文脈を見ることで主題化された文への選好が強まった）
- NTC 上と文脈を見せた場合のアノテーションで主題化されていない文が選ばれた，かつ文脈の有無による選好の変化の大きさが下位 25%に属する（文脈を見ることで主題化されていない文への選好が強まった）

実験では，文脈依存セットを用いる．

3　実験設定

3 種類の単方向言語モデルを用いて，人間と言語モデルの主題化に関する選好を比較する．

3.1　言語モデル

2 つの Transformer 言語モデル Trans-l（400M パラメータ），Trans-s（55M パラメータ）と，LSTM 言語モデル（55M パラメータ）を実験に用いた [20, 21]．言語モデルは新聞記事と Wikipedia からなる 300 万文章で学習した[8]．言語モデルへの入力は，JUMAN [22] で形態素に分割したのち，Unigram モデル [23] でサブワードに分割した[9]．

3.2　評価指標

2 節で作成した各インスタンス $(c_i, s_i^{\mathrm{TOP}}, s_i^{\mathrm{NOT}})$ について，人間と言語モデルの主題化率 r_i を以下のように計算する．

$$r_i = \frac{n(s_i^{\mathrm{TOP}})}{n(s_i^{\mathrm{TOP}}) + n(s_i^{\mathrm{NOT}})} \quad . \tag{1}$$

人間の選好については，$n(\cdot)$ としてクラウドソーシングで得られたどちらの文が自然かに関する票数を

8）　データサイズはトークナイズ前で 3.4GB である．また学習データは，2 節で作成したデータと重複しない

9）　sentencepiece [24] を用いた（character coverage=0.9995, vocab size=100000）

表 3　文脈依存セットでの人間と言語モデルの選択の傾向

モデル	文脈	主格 [17]			与格			対格		
		F1	ρ_r	ρ_Δ	F1	ρ_r	ρ_Δ	F1	ρ_r	ρ_Δ
TRANS-L	○	83.5 81.7	0.67 0.60	−0.12	88.4 89.9	0.68 −0.56	−0.19	86.2 86.9	0.62 −0.65	−0.01
TRANS-S	○	85.3 83.7	0.72 0.61	−0.07	88.2 86.9	0.71 −0.51	−0.16	82.3 80.8	0.53 −0.59	0.01
LSTM	○	81.9 82.3	0.69 0.62	−0.20	88.2 89.8	0.69 −0.50	−0.40	80.6 79.8	0.48 −0.58	−0.25
HUMAN	○	(100) 81.1	- -	-	(100) 19.8	- -	-	(100) 18.8	(1.0) -	-

用いる[10]．言語モデルの場合は，各文に対するパープレキシティを $n(\cdot)$ として用いる．r_i が大きいほど主題化された文が強く選好されることを意味する．文脈を提示した下で，人間と言語モデルの主題化率 r_i の相関 ρ_r を求め，人間が強く主題化を選好する事例において言語モデルも強く主題化を選好するか調査する．なお，言語的な規則・制約において判断が完全に定まる問題ではないため，2 値には要約せず，主題化率という連続値のまま扱う．

また，文脈 c_i のもとで得られた主題化率を r_i^{C}，文脈を考慮せずに得られた主題化率を r_i と区別し，各事例において，文脈が主題化の選択をどれほど促すか $\Delta = r_i^{\mathrm{C}} - r_i$ を計算した．Δ が大きいほど，文脈 c が，項の主題化を強く促したことを意味する．[11] 主題化の選択における先行文脈の影響の度合いが人間と言語モデルで近いかを調べるために，人間と言語モデルから求まる文脈の影響度 Δ の順位相関係数 ρ_Δ も報告する．また参考として，NTC 上での主題化の選択を正解と見做し，主題化する/しないの 2 値分類として扱った際の F1 値も報告する[12]．

4　文脈依存セットでの結果

2.3 節で作成した文脈依存セットにおける，人間と言語モデルの選択の傾向を表 3 に示す．文脈ありの設定では，人間と言語モデルの判断について，主格，与格，対格とも正の相関が観察された（$\rho_r \geq 0.48$）．一方，文脈の影響度の相関 ρ_Δ は，主格，与格，対格のいずれにおいても非常に弱く，主題化に対する文脈の影響について，人間と言語モデルで乖離があることが示唆された．

また，人間の F1 値は文脈の有無によって大きく変化しているのに対し，言語モデルの F1 値はほとんど変わらず，言語モデルが不当に文脈非依存な選択を行っていることが示唆された．特に与格や対格などコーパスと比較して人間の選択が著しく変わる設定において，なぜ言語モデルが文脈なしにコーパスと一貫した選択が行えているのかは興味深い．なお，文脈依存セットの作成では，人間とコーパスの判断が一致することを条件に課したため，文脈ありの設定で人間の F1 値は 100 となる．また文脈の有無で主題化における選好が大きく変わるデータを収集したため，文脈がないときに人間は積極的にコーパスと逆の選択をとる可能性があり，F1 値がチャンスレートを下回ることは不自然でない．なおデータセット全体で評価した場合でも，文脈の影響度の相関 ρ_Δ は依然として非常に弱く，一貫して言語モデルの人間と乖離した振る舞いが確認された（付録 A；表 4）．

5　おわりに

本研究では，談話の一側面である主題構造に焦点を当て，主題化の判断における人間と言語モデルの選好の比較を行った．実験では，人間が文脈に依存して判断を行う事例に対しても言語モデルは文脈非依存な判断を行うこと，主題化に対する文脈の影響について人間と言語モデルで乖離があることを示した．この結果は，人間と同様の談話レベルの言語的知識を言語モデルに獲得させるためには，モデルのアーキテクチャや学習において更なる帰納バイアスが必要であることを示唆している．

10）「どちらでも良い」という票については，$s^{\mathrm{TOP}}, s^{\mathrm{NOT}}$ に 0.5 票ずつ分配した．

11）なお一連の計算において確率の足し算・引き算が発生しており，確率的な意味づけは困難になるが，比率などを用いた場合には今回の目的と反する性質を持つ値となり（例えば $\Delta = r_i^{\mathrm{C}}/r_i$ とした場合，1 票から 2 票への変化が 2 となり，7 票から 8 票の変化が 1.1 程度となるのは今回の文脈では不自然），この定式化に至っている．

12）人間・言語モデル共に，主題化率 r_i が 0.5 を上回る場合には s^{TOP} を選択したとみなし，そうでない場合は s^{NOT} を選択したとみなした．

謝辞

本研究は JST CREST JPMJCR20D2 の助成を受けたものです．また，データセットの作成にあたり毎日新聞記事データ集 1995 年版のクラウドソーシングでの使用を許可してくださった毎日新聞社に感謝致します．

参考文献

[1] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 521–535, 2016.

[2] Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. **Cognitive Science**, Vol. 41, No. 5, pp. 1202–1241, 2017.

[3] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In **Proceedings of EMNLP**, pp. 1192–1202, 2018.

[4] Yoav Goldberg. Assessing bert's syntactic abilities. **arXiv preprint arXiv:1901.05287**, 2019.

[5] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 625–641, 2019.

[6] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[7] Jiwei Li and Dan Jurafsky. Neural net models of open-domain discourse coherence. In **Proceedings of EMNLP**, pp. 198–209, September 2017.

[8] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. Do massively pretrained language models make better storytellers? In **Proceedings of CoNLL**, pp. 843–861, 2019.

[9] Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. Cohesion in english. **Longman**, 1976.

[10] Michael Halliday, Christian MIM Matthiessen, and Christian Matthiessen. **An Introduction to Functional Grammar**. Routledge, 2014.

[11] Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse. **Computational Linguistics**, Vol. 21, No. 2, pp. 203–225, 1995.

[12] Marilyn Walker, Masayo Iida, and Sharon Cote. Japanese discourse and the process of centering. **Computational Linguistics**, Vol. 20, No. 2, pp. 193–231, 1994.

[13] 松下大三郎. 標準日本口語法. 中文館書店, 1930.

[14] 孟玲秀. 『日本語教育における「は」と「が」の教授法』 ——中国人学習者に対する日本語教育の場合——. 2004.

[15] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In **Proceedings of the linguistic annotation workshop**, pp. 132–139, 2007.

[16] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with mace. In **Proceedings of NAACL-HLT**, pp. 1120–1130, 2013.

[17] Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Ryoko Tokuhisa, and Kentaro Inui. Topicalization in language models: A case study on Japanese. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 851–862, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[18] 野田尚史. 「は」と「が」. くろしお出版, 1996.

[19] 日本語記述文法研究会. 現代日本語文法 5 とりたて・主題. くろしお出版, 2009.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of NeurIPS**, pp. 5998–6008, 2017.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[22] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In **Proceedings of LREC**, pp. 1344–1347, 2006.

[23] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In **Proceedings of ACL**, pp. 66–75, 2018.

[24] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of EMNLP**, pp. 66–71, 2018.

表 4　データセット全体での人間と言語モデルの選択の傾向

モデル	文脈	主格 [17]			与格			対格		
		F1	ρ_r	ρ_Δ	F1	ρ_r	ρ_Δ	F1	ρ_r	ρ_Δ
TRANS-L	○	88.1	0.80	−0.04	89.3	0.26	−0.12	83.6	0.25	0.03
		87.5	0.78		88.5	−0.06		81.6	−0.06	
TRANS-S	○	87.7	0.80	−0.02	85.0	0.26	−0.10	84.9	0.21	0.04
		87.3	0.78		83.6	−0.08		82.9	−0.01	
LSTM	○	85.2	0.79	−0.01	81.9	0.24	−0.11	79.3	0.21	−0.05
		85.3	0.78		82.3	−0.03		78.9	−0.02	
HUMAN	○	89.7	-	-	62.5	-	-	59.6	-	-
		89.1	-		43.4	-		49.0	-	

A　データセット全体での結果

表 4 に 2 節で作成したデータセット全体での人間と言語モデルの選択の傾向を示す．データセット全体においても主格，与格，対格いずれの文脈の影響度合いの相関 ρ_Δ はほとんどなく，4 節の結果と一貫して，人間と言語モデルの文脈の考慮の仕方には乖離があることが確認された．

F1 値については，主格における主題化の判断では，人間も言語モデルも 90 ポイント前後と高い値を示した．また，データセット全体では人間も言語モデルも文脈の有無による F1 値の変化が小さく，主格における主題化の判断，すなわち「は」と「が」の選択は文内の情報のみで十分可能であることが示唆された．一方で，与格と対格における主題化の判断では，言語モデルは 80 から 90 ポイント前後の値となっているのに対し，人間は 50 から 60 ポイント前後の値となっていた．この値は 2 値分類のチャンスレートに近いことから，人間にとって与格と対格における主題化判断は一方に選好が偏るものではないことが考えられる．しかしながら，文脈を見ることでコーパスと一貫した判断がわずかではあるができるようになるという点で，人間は与格や対格における主題化の判断を文脈に依存して行っていることが改めて確認された．人間と比べて言語モデルは F1 値で高いスコアとなっているが，依然として文脈に依存した判断を行っていないことから，過剰にコーパスにフィットしていたり，N-gram レベルの情報で判断をしていたりする可能性が考えられる．人間にとっては困難な判断を言語モデルがどんな情報を頼りに行っているのかということについては今後さらに調査したい．