日本語の倒置が大規模言語モデルの読解性能に与える影響

津田 純花 [1]　矢野 一樹 [1]　赤間 怜奈 [1,2,3]　鈴木 潤 [1,3,4]
[1] 東北大学　[2] 国立国語研究所　[3] 理化学研究所　[4] 国立情報科学研究所 LLMC
is-failab-research@grp.tohoku.ac.jp

概要

日本語は語順の自由度が高く，多様な修辞技法が存在するが，大規模言語モデルがそれらをどの程度理解し，処理できるかは，十分には検討されていない．本研究では，語順を変更する「倒置」に着目し，日本語機械読解タスクを用いて，倒置が大規模言語モデルの読解性能に与える影響を定量的に評価した．その結果，多くのモデルにおいて，倒置による性能の低下が確認された．さらに，トークンごとのサプライザルを分析することにより，倒置構造がモデル内部の処理負荷を増大させることを示した．

1　はじめに

自然言語には，比喩や倒置をはじめとする多様な修辞技法が存在している [1]．修辞技法は，重要な情報の強調や，印象の変化などを通じて，より効果的な表現を生み出すことができる一方で，解釈のずれや曖昧さを引き起こし，書き手の意図が十分に伝達されない可能性がある．したがって，大規模言語モデル（LLM）においても，修辞技法を含む文を適切に処理することは重要な課題であるが，比喩や含意といった修辞技法について，LLM は十分に理解できていないと報告されている [2, 3]．

本研究では，修辞技法の中でも「倒置」に着目する．日本語は自然言語の中でも，語順の自由度が高い言語であることが知られている [4, 5]．語順を意図的に変更する倒置は，情報の強調や補足の目的で用いられる [6]．文学作品や記事の見出し，日常会話など，幅広い場面で観察され，日本語において自然に用いられる修辞技法の一つである．

語順の変更が LLM の性能に与える影響については，これまでにも検討が行われてきた [7, 8, 9]．しかし，その多くは英語を対象とし，ランダムな語順操作に基づくものであり，倒置のような自然な語順変更とは異なる性質をもつと考えられる．

本研究では，倒置を「通常は述語よりも前に現れる主語や修飾語などを，意図的に述語の後に配置する語順変更」と定義する．異なる特性をもつ 2 種のデータセットを用いて，倒置が LLM の読解性能に与える影響を定量的に測定し，分析を行った．

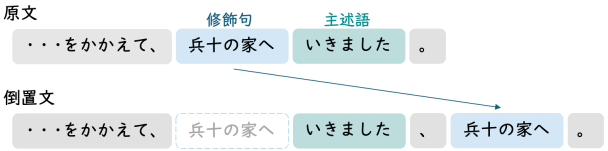


図 1　修飾語倒置の生成例．修飾句を元の位置から削除し，文末に移動することで倒置文を生成した．

実験の結果，倒置文を用いた場合では，LLM の読解性能が原文と比べて，低下する傾向がみられた．さらに，モデルの内部的な処理負荷を反映する指標としてサプライザルを分析したところ，倒置文においてサプライザルの上昇が確認された．意図的に語順を変更する倒置が，LLM の読解に対して一定の負荷を与えることが示された．

2　関連研究

既存研究では，語順の変更が LLM の性能に与える影響についてさまざまな観点から検討されてきた．語順をランダムにシャッフルしたコーパスを用いて事前学習した場合や，入力文の語順を大きく崩した場合であっても，多くのタスクにおいて性能の低下が限定的であることが報告されている [7, 8]．一方で，語順による影響はタスクに依存することが指摘されており，読解や推論のように文中の意味的対応関係が重要となるタスクでは，語順の変更がモデルの読解性能に大きな影響を与えることがあると報告されている [9]．

本研究では，意図的に語順を変更する「倒置」という修辞技法に着目し，倒置が LLM の読解性能に与える影響を明らかにする．

3 倒置文の生成

3.1 倒置文生成手法

日本語文に対する倒置文を自動生成するため，spaCy の日本語モデル[1] を用いた依存構造解析に基づくルールベース手法を採用した．各文に対して構造解析を行い，主語，目的語，および修飾語に対応する句を検出した．倒置可能な句が存在する場合に，抽出した句を元の位置から除去し，文末の終止符の直前に移動させることで倒置文を生成した．図 1 に倒置文の生成手法を示す．

3.2 倒置文の品質チェック

自動生成された倒置文の妥当性を評価するため，Yahoo!クラウドソーシングを用いた人手評価を実施した．評価対象は，汎用文および使用するデータセット（JSQuAD，JFairytaleQA[2]）より抽出した文であり，内訳は汎用文が 62 文，JSQuAD および JFairytaleQA から抽出した文が 15 文ずつの計 92 文である．各文につき，3 名のワーカーに評価を依頼した．評価項目は「自然さ」「形式」「意味の保持」の 3 項目とし，いずれも 2 段階での判定を求めた．また，回答の信頼性を確保するため，チェック設問を設け，所定の基準を満たさないワーカーの回答は除外した．最終的な評価結果は，3 名のワーカーの回答による多数決に基づいて決定し，自動生成された倒置文の品質を分析した．

人手評価の結果を，表 1 および表 2 に示す．表 1 より，汎用文においては，自然さ・形式・意味の保持のいずれの観点においても高い評価が得られており，本研究で用いる倒置生成手法が，基本的な文構造をもつ文に対して妥当であることが確認された．さらに，汎用文のうち文長が短い文（30 文字以下）に限定すると，評価スコアはさらに改善し，すべての評価項目において 90% 以上を記録した．

一方，JSQuAD より抽出した文では，自然さおよび意味の保持の評価が相対的に低い結果となった．これは，文が説明的かつ複雑であることや，文長が比較的長いことが影響した可能性がある．また，JFairytaleQA より抽出した文では，すべての評価項目において JSQuAD より抽出した文を上回る結果が

1） spaCy の日本語モデル ja_core_news_md を用いた．https://spacy.io/models/ja

2） JFairytaleQA は，新たに作成した日本語物語データセットであり，詳細は第 4.1 節に記載する．

表 1 文種別の倒置文評価結果．

	文数	自然さ (%)	形式 (%)	意味保持 (%)
汎用文	62	75.81	82.26	91.94
JSQuAD	15	53.33	73.33	66.67
JFairytaleQA	15	66.67	80.00	86.67

表 2 倒置句の種類別の評価結果．

	文数	自然さ (%)	形式 (%)	意味保持 (%)
主語	30	70.00	70.00	90.00
目的語	31	64.52	87.10	83.87
修飾語	31	**77.42**	83.87	87.10

得られており，文学的文脈において倒置が比較的自然に解釈されやすい可能性が示唆される．

次に，倒置句の種類別の結果に着目する．表 2 より，修飾語を含む句の倒置が最も自然さの評価が高く，形式および意味の保持においても安定した結果となった．そのため，本研究では倒置の対象を修飾語に限定することとした．

4 実験設定

4.1 データセット

倒置が LLM の読解性能に与える影響を検討するため，JSQuAD および本研究で新たに構築した JFairytaleQA を使用した．倒置による影響が読解対象の性質に依存するかを検討するため，特性の異なる 2 種のデータセットを用いた．

JSQuAD 機械読解データセットである SQuAD の日本語版で，Wikipedia の記事を基に作成された段落と，それに関する質問および正解から構成される [10, 11]．日本語機械読解タスクとして広く利用されていることから，倒置が読解性能に与える影響を評価するためのデータセットとして採用した．

JFairytaleQA 日本語物語文とそれに関する質問および正解から構成される日本語機械読解データセットである．FairytaleQA [12] を参考に，本研究で新たに構築した[3]．JFairytaleQA に含まれる日本語物語は，青空文庫[4]に収録されているもののうち，児童向け小説・物語を表す日本十進分類法（NDC）の分類コード K913 が付与された作品を対象に収集した．難易度の高い作品を除外するため，jreadability [13] を用いて各作品の可読性指標を算出し，学習者レベル lower intermediate より上（スコアが 3.5 未満）の読解能力を必要とする作品を対象外

3） 現在データ規模を拡大中，完成後公開予定．

4） https://www.aozora.gr.jp/

表 3 JSQuAD における EM および F1．背景色は，倒置によりスコアが低下した箇所を示す．

モデル	EM			F1		
	原文	倒置文	差分	原文	倒置文	差分
llm-jp	0.749	0.735	− 0.014	0.863	0.851	− 0.011
Qwen	0.701	0.686	− 0.016	0.854	0.852	− 0.002
rinna	0.542	0.515	− 0.027	0.728	0.705	− 0.023
sarashina	0.748	0.722	− 0.026	0.873	0.860	− 0.013
Swallow	0.822	0.790	− 0.032	0.917	0.900	− 0.018

表 4 JFairytaleQA における EM および F1．背景色は，倒置によりスコアが低下した箇所を示す．

モデル	EM			F1		
	原文	倒置文	差分	原文	倒置文	差分
llm-jp	0.081	0.073	− 0.009	0.369	0.344	− 0.025
Qwen	0.094	0.085	− 0.009	0.453	0.423	− 0.029
rinna	0.013	0.021	+ 0.009	0.281	0.276	− 0.004
sarashina	0.107	0.107	± 0.000	0.500	0.480	− 0.020
Swallow	0.162	0.145	− 0.017	0.535	0.527	− 0.009

とした．教科書掲載歴のある作品やアクセスランキング上位の作品など，日本語話者にとって特に馴染みのある児童文学作品を中心に，合計 18 作品を選定した．データセット化に際し，旧字旧仮名による倒置生成への影響を排除するとともに，データセットとしての汎用性を高めるため，文字遣い種別は新字新仮名に統一した．

質問は，文章中から直接解答を抽出できる explicit 型と，文脈理解や推論を要する implicit 型を含む．また，FairytaleQA に従い，登場人物，状況設定，行動，心情，因果関係，結果，推測の 7 つのカテゴリーに分類される．質問および解答の作成は，クイズ作成経験のある専門家に依頼し，一貫した基準に基づく品質を確保した．物語文は，叙述的な展開や語りの構造が特徴的であり，説明文主体の JSQuAD とは異なる特性をもつ．そのため，倒置が LLM の性能に与える影響の包括的な分析が可能となる．

4.2 評価対象とした言語モデル

日本語機械読解タスクにおける倒置の影響を評価するために，本実験では，日本語に特化して訓練されたモデル，あるいは日本語において高い性能を示すことが報告されている 5 種のモデル（llm-jp-3-1.8b-instruct[5] [14], Qwen2.5-3B-Instruct[6] [15], japanese-gpt-neox-3.6b-instruction-sft-v2 (rinna)[7] [16], sarashina2.2-3b-instruct-v0.1[8], Llama-3-Swallow-8B-Instruct-v0.1[9] [17]）を採用した．

5） https://huggingface.co/llm-jp/llm-jp-3-1.8b-instruct
6） https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
7） https://huggingface.co/rinna/japanese-gpt-neox-3.6b-instruction-sft-v2
8） https://huggingface.co/sbintuitions/sarashina2.2-3b-instruct-v0.1
9） https://huggingface.co/tokyotech-llm/Llama-3-Swallow-8B-Instruct-v0.1

4.3 評価指標

倒置が LLM の読解性能に与える影響を評価するため，評価指標として Exact match（EM）および F1 を用いた [11]．EM と F1 の 2 つの評価指標を用いることで，厳密な正解一致と部分的一致の両面から，倒置による読解性能への影響を検討した．F1 は，予測された解答と正解における適合率および再現率の調和平均であり，部分一致を考慮した指標である．形態素解析機による影響を除外するため，文字単位で F1 を算出した．

加えて，倒置がモデル内部の処理負荷に与える影響を分析するため，サプライザル [18] を算出した．サプライザルは，各トークンの負の対数尤度として定義され，値が大きいほど，モデルにとって予測困難であることを示す．すなわち，そのトークンが意外性の高い入力であることを意味する [19]．

5 実験結果

5.1 JSQuAD による評価

表 3 に JSQuAD における，原文と倒置文に対する各モデルの EM および F1 を示す．JSQuAD では，各サンプルの段落に対して倒置を施したデータセットを作成した．質問および正解は，原文と倒置文のデータセットで共通である．

表 3 の差分より，すべてのモデルにおいて，倒置文では EM および F1 が低下しており，倒置によって読解性能が影響を受けることが示された．この傾向は一貫して観察され，倒置が日本語機械読解タスクにおいて，負の影響を与えることが示唆される．

5.2 JFairytaleQA による評価

表 4 に，JFairytaleQA における原文と倒置文に対する各モデルの EM および F1 を示す．JFairytaleQA では，各サンプルの物語本文に対して倒置を施した

表 5　JSQuAD における平均文長別の F1 の差分．太字は各モデルにおいて，倒置による F1 の変化が最大の箇所を示す．

平均文長	F1 の差分		
	< 30 (n = 856)	30 – 60 (n = 2773)	> 60 (n = 791)
llm-jp	+ 0.000	– **0.018**	– 0.001
Qwen	– 0.000	– 0.005	+ **0.005**
rinna	– **0.026**	– 0.024	– 0.017
sarashina	– **0.017**	– 0.014	– 0.005
Swallow	– 0.011	– 0.019	– **0.020**

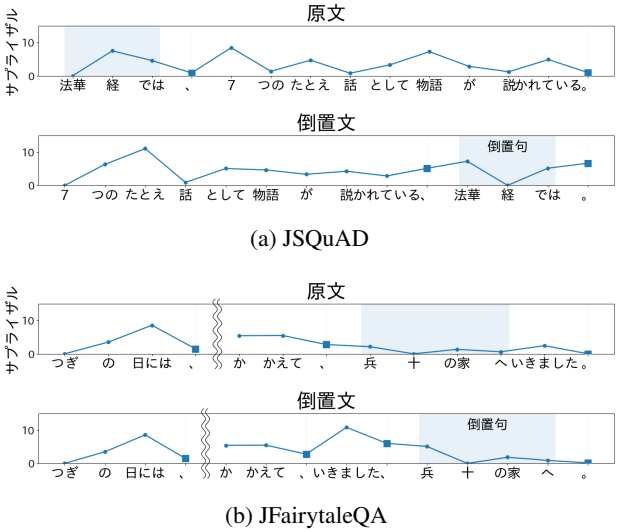
データセットを作成した．質問および正解は，原文と倒置文のデータセットで共通である．表 4 の差分より，多くのモデルで倒置文における EM および F1 の低下が確認された．JSQuAD のような説明文に限らず，叙述的な文脈をもつ物語文においても，倒置が読解性能に影響を及ぼすことが示された．

JFairytaleQA は新たに作成したデータセットであり，物語文における読解タスクの性質上，EM や F1 といった自動評価指標のみではモデルの出力を十分に評価できていない可能性が残る．自動評価指標の信頼性を担保するため，人手評価も実施した．特に，倒置によって正解表現の表出が変化した場合や，正解が一意に定まらない場合に配慮する必要がある．そのため，EM および F1 の結果が対照的であった rinna と Swallow の 2 種のモデルを対象とし，筆者による人手評価を行った．

その結果，人手評価においても，倒置文ではスコアが低下する傾向を確認できた．また，人手評価による傾向は，自動評価指標による傾向と一致しており，JFairytaleQA における EM および F1 の妥当性を確認できた（詳細は付録 A.2）．

5.3　定量分析：倒置における文長の影響

文の長さが倒置による読解性能の低下に影響を与えるかを分析するため，JSQuAD の段落の平均文長に基づいて F1 を算出した．表 5 に倒置文の F1 から原文の F1 を引いた差分を示す．倒置が読解性能に与える影響と文長との関係に一貫した傾向は見られず，モデルごとに異なる挙動が確認された．この結果は，倒置の影響の現れ方がモデルによって異なることに加え，自然な倒置が用いられている場合においても，読解性能が低下する可能性を示唆している．クラウドソーシングによる人手評価（第 3.2 節）で短い文ほど自然な倒置と受け取られやすいことが確認されているが，そのような場合においても倒置



(a) JSQuAD

(b) JFairytaleQA

図 2　原文および倒置文での各トークンにおけるサプライザルの比較．(a) は JSQuAD，(b) は JFairytaleQA より抜粋．倒置文では，倒置句の周辺でサプライザルが上昇．

により読解性能の低下が確認された．

5.4　定性分析：サプライザルの変化

図 2 に，トークン単位のサプライザルを観察した結果を示す．サプライザルの変化より，倒置がモデルにどのような影響を与えるかを分析した．JSQuAD および JFairytaleQA の両データセットの倒置文において，倒置句の周辺においてサプライザルの上昇がみられた．句読点でのサプライザルの上昇が顕著である例（図 2(a)）や，語順が変更された句の先頭トークンでサプライザルが上昇する例（図 2(b)）が確認された．これより，倒置は語順が変化した周辺において，モデル内部の処理負荷を局所的に増大させることが示唆される．

加えて，JSQuAD の段落における平均サプライザルを測定した．倒置文では原文と比較し，一貫して高い平均サプライザルを示した（詳細は付録 A.3）．

6　おわりに

本研究では，日本語の倒置が LLM の読解性能に与える影響について，日本語機械読解タスクである JSQuAD および JFairytaleQA を用いて検討した．その結果，日本語において意図的に語順を変更する倒置であっても，LLM の読解性能の低下を引き起こすことが確認された．本研究の結果は，修辞技法の一つである倒置を，LLM が十分に理解できていない可能性を示唆している．

今後は，JFairytaleQA の拡張とともに，倒置と LLM の読解性能の関係を体系的に検討していく．

謝辞

本研究を進めるにあたり，ご指導ならびにご助言を賜りました Tohoku NLP Group の皆様に感謝申し上げます．また，データセットの作成およびアノテーション作業にご協力いただいた皆様に感謝申し上げます．

本研究は，JSPS 科研費 JP25K21263, JST BOOST JPMJBY24A1, JST BOOST JPMJBS2421, JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research)，および，文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の助成を受けたものです．

参考文献

[1] Raymond W. Gibbs. **The Poetics of Mind: Figurative Thought, Language, and Understanding**. Cambridge University Press, 1994.

[2] Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4437–4452, 2022.

[3] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8690–8705, 2020.

[4] Masayoshi Shibatani. **The languages of Japan**. Cambridge University Press, 1990.

[5] Ryohei Sasano and Manabu Okumura. A corpus-based analysis of canonical word order of Japanese double object constructions. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2236–2244, 2016.

[6] 藤井洋子. 日本語文における語順の逆転 — 談話語用論的視点からの分析. 言語研究, Vol. 99, pp. 58–81, 1991.

[7] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2888–2913, 2021.

[8] Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. Bert & family eat word salad: Experiments with text understanding. In **Proceedings of the AAAI Conference on Artificial Intelligence**, pp. 12946–12954, 2021.

[9] Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 1145–1160, 2021.

[10] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2383–2392, 2016.

[11] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, 2022.

[12] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 447–460, 2022.

[13] Yoichiro Hasebe and Jae-Ho Lee. Introducing a readability evaluation system for japanese language education. In **Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese**, pp. 19–22, 2015.

[14] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. **arXiv preprint arXiv:2407.03963**, 2024.

[15] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, et al. Qwen2.5 technical report. **arXiv preprint arXiv:2412.15115**, 2024.

[16] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the japanese language. **arXiv preprint arXiv:2404.01657**, 2024.

[17] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. **arXiv preprint arXiv:2404.17790**, 2024.

[18] John Hale. A probabilistic earley parser as a psycholinguistic model. In **Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics**, pp. 159–166, 2001.

[19] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, pp. 1126–1177, 2008.

A　参考情報

A.1　Few-shot 数の検討

JFairytaleQA において，プロンプト中に与える few-shot 数が LLM の性能に与える影響を確認するため，Qwen および Swallow を対象に，few-shot 数を 0 から 5 まで変化させた予備実験を実施した．なお，本予備実験では，explicit な質問のみに限定し，少量のデータを対象とした．図 3 に示す通り，両モデルとも，few-shot 数を増加させることにより，性能の向上がみられ，徐々に緩やかになる傾向が見られた．その結果を踏まえ，4-shot を採用し，JFairytaleQA の評価を実施した．

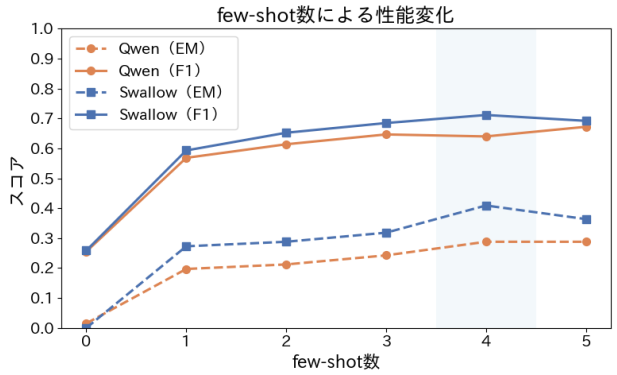


図 3　JFairytaleQA における few-shot 数の違いによる性能変化．Qwen および Swallow の両モデルにおいて，few-shot 数 4 前後で性能が最大となる傾向がみられた．

A.2　人手評価の結果

rinna と Swallow に対して実施した人手評価の結果を表 6 に示す．どちらのモデルにおいても，EM および F1 同様，倒置によるスコアの低下が確認できた．

表 6　JFairytaleQA における人手評価の結果．EM および F1 同様，倒置文でスコアが低下．

モデル	原文	倒置文
rinna	0.145	0.141
Swallow	0.675	0.632

A.3　サプライザル

JSQuAD の文脈における各モデルの平均サプライザルを表 7 に示す．段落ごとに 1 トークンあたりのサプライザルを算出し，それらを平均した値を平均サプライザルとして用いた．使用したすべてのモデルにおいて，サプライザルの上昇が確認され，各モデルにおいて，段落の 97%以上で倒置文の方が高い平均サプライザルを記録した．また，平均サプライザルの相対的な増加率に着目すると，倒置文では原文と比較して，約 13%から 18%の増加が確認された．

A.4　倒置の対象となった文の割合

使用したデータにおいて，倒置が施された文の割合を算出した．同一の文章が複数の質問に対応する場合があるため，文の重複を含め，算出した．その結果を表 8 に示す．JSQuAD では 90%，JFairytaleQA では 70% を超す文が倒置の対象となった．JFairytaleQA において，倒置の対象となる文の割合が相対的に低くなった要因としては，会話部に短い文が多く含まれていることが挙げられる．

表 7　JSQuAD の文脈における平均サプライザル．

モデル	平均サプライザル	
	原文	倒置文
llm-jp	3.247	3.670
Qwen	2.586	2.939
rinna	3.257	3.779
sarashina	2.949	3.496
Swallow	2.147	2.491

表 8　使用データにおける倒置対象文の割合．

データセット	全文数	倒置対象文数	割合 (%)
JSQuAD	13195	12085	91.59
JFairytaleQA	18260	13280	72.73

A.5　JFairytaleQA の具体例

JFairytaleQA で用いた質問および正解の例を表 9 に示す．各作品において，質問は 7 カテゴリーに対して明示的 / 暗黙的の区別を設けた 14 パターンのうち，推測・明示的を除く 13 パターンを作成した．

表 9　JFairytaleQA で用いた各カテゴリーの質問と正解例．

カテゴリー	例
登場人物	**Q**: 海で木の舟を作ったのは誰？ **A**: うさぎ
状況設定	**Q**: いちょうの木が立っているのはどこ？ **A**: 丘の上
行動	**Q**: 昨日、太郎は野原で何をした？ **A**: 凧を揚げた
心情	**Q**: 母さん狐は子狐の帰りをどんな気持ちで待っていた？ **A**: 心配しながら
因果関係	**Q**: 風が自分よりも壁の方が偉いと考えるのはなぜ？ **A**: 吹きとばすことはできないから
結果	**Q**: 三日降り続いた雨で、川の水の量はどうなっていた？ **A**: どっとましていました
推測	**Q**: 蟹の親子は、やまなしを見に行った後どうした？ **A**: 自分たちの住む穴に戻って寝た