Topicalization in Language Models: A Case Study on Japanese

Riki Fujihara[1] **Tatsuki Kuribayashi**[1,2] **Kaori Abe**[1]
Ryoko Tokuhisa[1] **Kentaro Inui**[1,3]
[1]Tohoku University [2]Langsmith Inc. [3]RIKEN
riki.fujihara.s4@dc.tohoku.ac.jp
{kuribayashi, abe-k, tokuhisa, inui}@tohoku.ac.jp

Abstract

Humans use different wordings depending on the context to facilitate efficient communication. For example, instead of completely new information, information related to the preceding context is typically placed at the sentence-initial position. In this study, we analyze whether neural language models (LMs) can capture such discourse-level preferences in text generation. Specifically, we focus on a particular aspect of discourse, namely the topic-comment structure. To analyze the linguistic knowledge of LMs separately, we chose the Japanese language, a topic-prominent language, for designing probing tasks, and we created human topicalization judgment data by crowdsourcing. Our experimental results suggest that LMs have different generalizations from humans; LMs exhibited less context-dependent behaviors toward topicalization judgment. These results highlight the need for the additional inductive biases to guide LMs to achieve successful discourse-level generalization.

1 Introduction

Building on the current success of neural language models (LMs) in the field of natural language processing (NLP), much work has been conducted to test their linguistic knowledge, typically, syntactic generalizations in LMs (Linzen et al., 2016; Lau et al., 2017; Marvin and Linzen, 2018; Goldberg, 2019; Warstadt et al., 2019, 2020). Although discourse is also an essential aspect of language production along with syntactic constructions, discourse-level knowledge in LMs has been less explored or has been typically analyzed at the coarse level, for example, analyzing their sentence ordering abilities (Li and Jurafsky, 2017; See et al., 2019).

As one step toward understanding the fine-grained, discourse-level knowledge in LMs, this study explores the generalization performance of

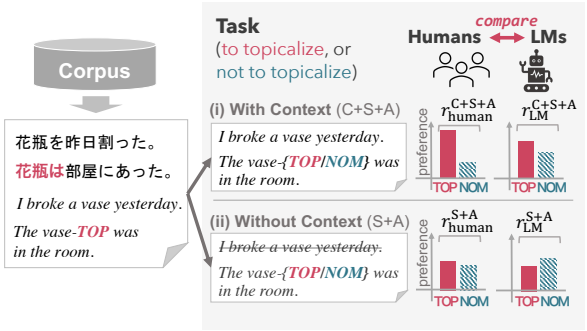

Figure 1: Comparing the context-dependent preferences of language models (LMs) and humans for topicalization.

LMs in a particular aspect of discourse, the topic-comment structure (i.e., the thematic structure). The topic-comment structure is an essential part of language production (Grosz et al., 1995; Halliday et al., 2014; Hajicová and Mírovskỳ, 2018).

For example, speakers use the following sentences in different contexts:

(1) a. In Japan, my father bought the vase last year.
b. The vase was the one my father bought in Japan last year.

These sentences differ in terms of topic-comment structure (the topic is underlined). For example, Sentence (1b) is more suited than (1a), for example, as a continuation of the sentence *I broke a vase yesterday.* This study probes whether LMs can make such context-dependent, human-like paradigmatic choices at the discourse level.

It is worth noting that the (non-)human-like linguistic generalization ability of computational models has long been examined in the cognitive science of language (Rumelhart and McClelland, 1985; Hu et al., 2020). Furthermore, from an engineering point of view, whether LMs have correct preferences for topicalization is an important perspective for validating the use of LMs for automatic assessment of text quality.

Proceedings of the 29th International Conference on Computational Linguistics, pages 851–862
October 12–17, 2022.

To test the cognitive plausibility of LMs' behaviors toward topic-comment structure, we used a topic-prominent language (TPL), where the topic-comment structure is explicitly realized. The use of a TPL facilitated the careful creation of text pairs for analyzing humans and LMs behaviors. We first annotated and analyzed the human preferences for topicalization using crowdsourcing (Sections 3 and 4).

In our experiments, we compared LMs with humans with respect to their context-dependent preferences for topicalization (Section 5; Figure 1). Our experimental results indicated that LMs do not show human-like, context-dependent behaviors (Section 6). Further analysis showed that LMs perform topicalization judgment based on context-independent spurious biases regarding whether a particular noun phrase is likely to be topicalized. This result reveals that compared to humans, LMs perform different generalizations toward topicalization judgment; hence, additional inductive biases are required to achieve human-like generalizations. Our dataset and code are publicly available. [1]

2 Background

Topic of a sentence. The topic of a sentence represents the concern of the message; what the speaker is going on to say. In the literature, the topic typically corresponds to the sentence-initial element (Halliday et al., 2014; Vallduví, 1990).[2] For example, the topic of Example (1a) (Section 1) is *Japan*, while that of Example (1b) is *vase*.

The term **topicalization** means to mark a particular element in a sentence as the topic of that sentence. Topicalization is realized typically by the word order as shown in Example (1), and different languages sometimes have different devices (e.g., particles and intonation) to mark a topic (Section 3.1).

The non-topic part of a sentence is called a comment. The topic and comment are often studied as a topic-comment structure or a theme-rheme structure in linguistics (Grosz et al., 1995; Halliday et al., 2014; Hajicová and Mírovskỳ, 2018). These are distinguished from the grammatical and logical structures of a sentence; for example, the topic of Example (1a) is *Japan*, while the grammatical subject is *my father*. Note that a sentence may not always have a topic (e.g., The sentence "There is a pen." does not introduce a topical theme).

Topicalization and discourse. In text production, the topic of a sentence plays an important role as it indicates the concern of the message and controls word ordering (Example (1)). In general, the more salient a particular concept is in a context, the more likely it is to be topicalized (Halliday et al., 2014; Miltsakaki, 1999). This property is closely related to the **centering** in discourse (Chafe, 1976; Givón, 1983; Grosz et al., 1995). In this theory, the topic of a sentence has a higher centering-forward (Cf) level according to the grammatical ranking, and such a higher-ranked entity is expected to have a high Cf rank again in the next utterance (i.e., the topic is expected to continue across successive utterances).

In this work, we determine whether LMs have human-like preferences for judging which elements in a sentence should be a topic (i.e., topicalize) when conveying meaning. Considering the context-dependent nature of topicalization, analysis of the preference of LMs in topicalization can provide insight to examine whether LMs capture the contextual flow of text at the inter-sentential, and discourse levels.

Linguistic probes. NLP researchers have inspected the inner workings and/or behaviors of black-box neural models to explore whether they actually understand the language. The focus of such probing analyses ranges from, for example, syntactic knowledge (Marvin and Linzen, 2018; Hu et al., 2020; Hewitt and Manning, 2019) to common sense knowledge (Lin et al., 2020; Zhou et al., 2020). Some analyses have also targeted discourse-level phenomena such as coreference (Sorodoc et al., 2020; Upadhye et al., 2020) and discourse structures (Pandia et al., 2021; Kurfalı and Östling, 2021; Koto et al., 2021). These existing studies and our present work are complementary in covering a wide variety of discourse phenomena.

To probe the linguistic knowledge of LMs, researchers have typically used minimally different text pairs (MDTPs) that differ *only* in a certain linguistic aspect and analyzed their probabilities computed by LMs (Marvin and Linzen, 2018; Gauthier et al.; Hu et al., 2020). This experimental design has advantages: for example, it enables researchers

[1] https://github.com/rk-fujifuji/lm_topicalization

[2] Strictly speaking, there are several definitions of the topic (e.g., textual, interpersonal, and topical); in this study, we focuses on the topical theme introduced in Halliday et al. (2014).

to analyze the model behaviors directly, without designing additional classifiers (Alain and Bengio, 2017; Pimentel et al., 2020). We also probe LMs using MDTPs that differ *only* in the topic-comment structure.

3 Dataset: collecting human preferences

3.1 Using topic-prominent language

In English, creating the MDTPs differing only in the topic-comment structure is prohibitively difficult because the change of topicalized elements in a sentence accompanies a drastic change in the sentence structure and syntactic complexities (Example (1)). Thus, simply analyzing the preferences of LMs between sentences with different topic-comment structures in English may lead to confusing conclusions about which linguistic perspective LMs are actually sensitive to.

In contrast, in Japanese, a TPL, creating a set of MDTPs differing only in the topic-comment structure is possible. For example, the following two sentences in Japanese, one of the TPLs, differ *only* in the topic of the sentence:

(2) a. *Kabin-wa heya-ni at-ta.*
Vase-TOP room-DAT exist-PAST.
The vase was in the room.

b. *Kabin-ga heya-ni at-ta.*
Vase-NOM room-DAT exist-PAST.
There was a vase in the room.

In Japanese, the postpositional particle *wa* (TOP) is used as the *topic marker* for indicating the topic of a sentence as in Example (2a) (Teruya, 2004, 2007; Kuno, 1973; Noda, 1996).[3] The difference between *Kabin-wa* (*Vase*-TOP) and *Kabin-ga* (*Vase*-NOM) in Example (2) is whether the element (*Kabin*; *vase*) is marked as a topic or not.

Note that Japanese is an agglutinative language, in which the functional information (e.g., grammatical case) of an element is realized by postpositional particles. When a particular element is marked by *wa* (TOP), some originally used particles (e.g., *ga*; NOM) are omitted. That is, *Kabin-wa* (*Vase*-TOP) in Example (2a) plays the roles of the grammatical subject (NOM) and the topic of a sentence (TOP) both, but only *wa* is attached.

Example (2a), in which the grammatical subject (*kabin*; *vase*) is topicalized, should be preferred to Example (2b) in the context of talking about the vase. We analyzed whether LMs have such context-dependent preferences for using topic markers. Such a paradigmatic choice of topicalization is not determined by strict rules. Rather, both Example (2a) and (2b) are usually acceptable, but either sentence is sometimes *more natural* than the other, depending on the context. We created a dataset of such *degree*s of human preference for topicalization.

Note that grammatical topic (TOP) is assumed to have a higher Cf ranking than the subject (NOM) in Japanese centering theory (Walker et al., 1994); this task of context-dependently selecting TOP or NOM could be viewed as a task of estimating Cf from the perspective of centering theory.

3.2 Annotation task

Data preparation. We focused on the Japanese language as a representative of TPLs. Specifically, we analyzed the preference for topicalization of nominative arguments in Japanese sentences. We used the NAIST Text Corpus (NTC; Iida et al., 2007), which is commonly used for analyzing linguistic phenomena in Japanese. We collected nominative arguments and their belonging sentences that satisfied all of the following criteria from the NTC:

- An argument has a nominative relation to the verb that is closest to the end of a sentence, regarding the predicate-argument structure annotation.[4]
- An argument accompanies the topic marker (TOP) or nominative particle (NOM).
- An argument appears in the second, third, or fourth sentence in a paragraph.

Using these criteria, we collected the data for annotation $\mathcal{D} = \{(c, s, a)_d\}_{d=1}^{|\mathcal{D}|}$, where c is the inter-sentential context (preceding sentences within the same document), s is the intra-sentential context, and a is the nominative argument that satisfies the aforementioned criteria. This process yielded 1,661 data points, where 939 instances originally have the TOP particle for nominative argument a, and 722 instances have NOM. The examples are listed in Table 1.

[3] Strictly speaking, the contrast between *wa* (TOP) and *ga* (NOM) is not only about the topicalization, and has long been discussed in Japanese linguistics. We did not intend to claim that all the *wa* work as a topic marker; instead, we carefully selected the data points where *wa* is used as a topic marker (Section 3).

[4] Predicate-argument structure annotation in NTC is used.

Text		Human preferences		LM preferences	
Context c	Sentence s and targeted nominative argument a (underlined)	$r_{\text{human}}^{\text{C+S+A}}$	$r_{\text{human}}^{\text{S+A}}$	$r_{\text{LM}}^{\text{C+S+A}}$	$r_{\text{LM}}^{\text{S+A}}$
We consume large amounts of energy daily.	*Economic growth*-{TOP/NOM} *is closely related to energy use.*	0.83	1.00	0.66	0.64
It is said that a good start determines victory in yacht racing.	*A strong wind*-{TOP/NOM} *of 15 knots was blown in the sea.*	0.00	0.00	0.31	0.25
The seventh is the day of the "Seven Herbs of Spring" to pray for good health. At the Hanshin Department Store in Kita, Osaka, a free service of Nanakusagayu was offered from 8:00 a.m. Office workers, young women, and junior high school students on their way home from early morning kendo practice enjoyed its taste.	*The 500 meals*-{TOP/NOM} *prepared were gone in about an hour.*	1.00	0.50	0.58	0.58

Table 1: Examples of the instances $(c, s, a)_d$ and preferences of humans and LMs (TRANS-L) for topicalization. The human preference scores r_{human} are introduced in Section 3.2. The LMs preference scores r_{LM} are introduced in Section 5.1. The example texts are the English-translated versions of the original texts (Mainichi Shimbun article data 1995 version).

Annotation task. We collected the topicalization preferences using crowdsourcing. Specifically, for each instance $(c, s, a)_d$, we first masked the particle following the nominative argument a (*kabin*; *vase*) as follows:

(3) a. *Kabin-__ heya-ni at-ta.*
Vase-__ room-DAT exist-PAST.

Then, annotators were asked which postpositional particle TOP (*wa*; to topicalize) or NOM (*ga*; not to topicalize) is more natural to complete the blank __. The option of "both okay" was also available.

For each instance $(c, s, a)_d$, from four to eight subjects annotated the preference. Finally, for each instance $(c, s, a)_d$, we calculated **topicalization ratio** as follows:

$$r_{\text{human}} = \frac{n_{\text{TOP}}}{n_{\text{TOP}} + n_{\text{NOM}}} \quad , \tag{1}$$

where n_{TOP} and n_{NOM} are the number of votes for completing the blank (e.g., *Kabin-__*) with TOP and NOM, respectively. For "both okay," we considered TOP and NOM to have 0.5 votes for each.

Context ablation. To facilitate analyzing the context-dependent characteristics of the topicalization, annotators solved the task under three different conditions. Table 2 shows each condition with an example. Here, ✓ in the **context** column indicates that the inter-sentential context is shown to annotators, and ✓ in the **sentence** column indicates whether the intra-sentential context is shown. When the intra-sentential context is not shown, the nominal constituent alone is provided.

Subsequently, three variants of the topicalization ratio r_{human} for each instance $(c, s, a)_d$ were obtained under different ablation settings: (i) the ratio $r_{\text{human}}^{\text{C+S+A}}$ obtained with the C+S+A setting, (ii) $r_{\text{human}}^{\text{S+A}}$ with the S+A setting, and (iii) $r_{\text{human}}^{\text{A}}$ with the A setting. The examples of scores are listed in Table 1. In Section 5, we observed the preferences of LMs for topicalization and compared them with those of humans.

Intended use of the annotations. We used the annotations obtained in the C+S+A and S+A settings (with relatively high agreement) for our main experiments (Section 5). The data of the A settings were used in our additional analyses (Section 6) to test whether the LMs also exhibit such a human-like difficulty in this setting. We observed some interesting trends: LMs exhibited unreasonably good performance of topicalization prediction in the context-independent, A settings.

3.3 Crowdsourcing

Worker selection. We used crowdsourcing[5] to access Japanese subjects. Crowd workers solved the task of selecting TOP or NOM. To qualify the motivated crowd workers, we first created trial tasks, in which each worker answered 10 questions by selecting TOP or NOM. For this purpose, we framed the validation questions in advance, where

[5] https://crowdsourcing.yahoo.co.jp/

	context c	sent. s	arg. a	question: TOP or NOM for __?
C+S+A	✓	✓	✓	*Kabin-wo kinou wat-ta. Kabin-__ heya-ni at-ta.* *I broke a vase yesterday. {The vase was}/{There was a vase} in the room.*
S+A		✓	✓	*Kabin-__ heya-ni at-ta.* *{The vase was}/{There was a vase} in the room.*
A			✓	*Kabin-__* *Vase*

Table 2: Settings for solving the topicalization judgment tasks. The Japanese example sentence "*Kabin-wo kinou wat-ta. Kabin-__ heya-ni at-ta.*" means "*I broke a vase yesterday. {The vase was}/{There was a vase} in the room.*" in English. The question is whether to topicalize the word "vase" in the second sentence.

Setting	#labels	class distribution		
		TOP	Both okay	NOM
C+S+A	8,039	55.3%	1.49%	43.2%
S+A	8,094	52.9%	3.47%	43.6%
A	7,621	1.86%	96.7 %	1.42 %

Table 3: Statistics of the whole dataset. The #labels denote the number of workers who annotated the labels remaining after the post-processing. The class distribution. columns denote the percentages of TOP, "Both okay", and NOM in the answers.

our preferences were in agreement; one of the 10 questions is a validation example. Then, we listed the motivated workers who can answer the validation question correctly. At this stage, 164 workers were selected in the C+S+A and S+A settings, and 153 workers were selected in the A setting. [6]

Annotation. For each instance $(c, s, a)_d$ of the 1,661 data points collected in Section 3.2, eight of the motivated workers annotated the preferences. Each worker solved at least ten instances. The same worker is not annotated to the same instance in different settings. After the whole annotation process, we performed statistical post-processing to exclude workers in the bottom 30% of competence using MACE (Hovy et al., 2013). We removed the data points annotated by fewer than four qualified workers.

Finally, 23,745 decisions selecting TOP or NOM for 1,355 nominative arguments were collected, where 758 instances originally have the TOP particle for nominative argument a, and 597 instances have NOM.

3.4 Statistics

The number of human decisions and their class distribution is shown in Table 3. The length of the context c, sentence s, and nominative argument a was 96.3±53.7, 50.1±24.4, and 4.5±2.2 in the number of characters (mean±standard deviation), respectively.[7]

The annotation agreement was 0.689 and 0.699 for Krippendorff's alpha for the settings of C+S+A and S+A, respectively.[8] These values are above the minimum criteria for data reliability (0.667) (Krippendorff, 2004). In contrast, the agreement in the A settings (0.074) was far below the criteria. One plausible cause of such a low score is that the majority of annotators answered as "both okay" in these settings. This skewed the class distribution, and the alpha value is affected by class imbalances (Jeni et al., 2013). Since many workers answered "both okay," we tentatively conclude that making topicalization judgments in this setting is difficult for Japanese speakers.

4 Data analysis: Context effect in topicalization

Before our experiments, we preliminarily observed the characteristics of the collected data. If all the annotated preference remains unchanged regardless of the inter-sentential context, our data are not suitable for analyzing the discourse-level behaviors in LMs. Our analysis declines such a concern.

[6] We conducted crowdsourcing in two separate sessions; in the first session, we adopted the S+C+A and C+A settings, and in the second, we adopted the A setting. In each session, after the trial task, the worker's confidence level was calculated by MACE (Hovy et al., 2013). Then, the top 80% of workers (164 and 153 workers for the first and second session, respectively) were selected.

[7] The Japanese language has no explicit word boundary. As an approximation of word count, context c, sentence s, and nominative argument a have 56.4±31.7, 29.5±14.4, and 2.8±1.1 morphemes, respectively. We used a JUMAN dictionary for morphological analysis (Kawahara and Kurohashi, 2006).

[8] We used the weighted Krippendorff's alpha, where we assumed the distance scale as TOP ≺ "both okay" ≺ NOM.

#Mentions	#Arguments	$r^{\mathrm{C+S+A}}_{\mathrm{human}}$
0	1,082	0.48 ± 0.43
1	197	0.87 ± 0.25
2+	76	0.90 ± 0.22

Table 4: Frequently mentioned information in a context (i.e., higher #Mention) is likely to be topicalized (i.e., higher $r^{\mathrm{C+S+A}}_{\mathrm{human}}$). #Argument is the number of the corresponding data points. The mean and standard deviation of $r^{\mathrm{C+S+A}}_{\mathrm{human}}$ are presented.

Context effect on topicalization confidence. We first analyze the interaction between topicalization preference and context, regarding the linguistic theory that *already-mentioned, old information is more likely to be topicalized by the topic marker in Japanese* (Matsushita, 1930). We observed that the nominative argument frequently mentioned in its context tends to gain a higher topicalization ratio $r^{\mathrm{C+S+A}}_{\mathrm{human}}$ (Table 4). This supports that the created data reflect linguistically natural trends at the discourse level. Here, we used the co-reference annotation in the NTC to count how many times the same entity as the nominatives appeared in the preceding context.

Next, for each instance $(c,\ s,\ a)_d$, we quantified the context-dependent changes in topicalization preference as follows:

$$\Delta_{\mathrm{human}} = r^{\mathrm{C+S+A}}_{\mathrm{human}} - r^{\mathrm{S+A}}_{\mathrm{human}} \quad . \qquad (2)$$

Here, Δ denotes the change in the level of certainty in choosing TOP due to the presence of inter-sentential context.[9] Intuitively, a larger Δ indicates more votes on the topic marker when the context is available.

Figure 2 shows the distribution of the preference difference attributed to the inter-sentential context (Δ_{human}). We found that the preference changed depending on the presence of inter-sentential context (about 53% of data points have non-zero Δ_{human}).

Challenging set for probing LMs. The dataset has context-dependent nature, but there are also data points for which discourse-level context does not affect human behavior, we also created a **challenging** set. In this set, the context information is considered necessary for choosing TOP or NOM. Specifically, we selected the instances $(c,\ s,\ a)_d$ satisfying either of the following criteria:

- Original text and crowd workers demonstrated that a should be topicalized ($r^{\mathrm{C+S+A}}_{\mathrm{human}} > 0.5$), and inter-sentential context affected such a decision (Δ_{human} is in the top 25%).
- Original text and crowd workers demonstrated that a should not be topicalized ($r^{\mathrm{C+S+A}}_{\mathrm{human}} < 0.5$), and inter-sentential context affected such a decision (Δ_{human} is in the bottom 25%).

Finally, we obtained 311 instances, among which 209 instances originally had the TOP particle as the nominative argument a, and 102 instances had NOM. The statistics of the challenging set are shown in Appendix. We used this challenging set along with the whole dataset for our experiments to test the LMs (Section 5).

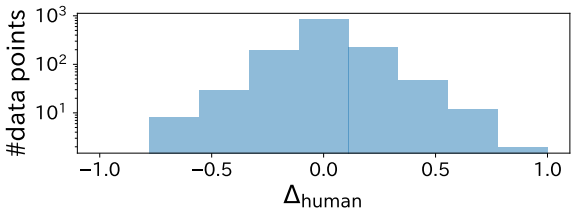

Figure 2: Histogram of the change in the topicalization ratio due to the presence of context (Δ_{human}).

5 Experiments: Comparing LMs with humans

Do LMs exhibit human-like topicalization preferences? To answer this question, we compared contrast the preferences of LMs and humans.

5.1 Experimental settings

Language models. We tested three variants of left-to-right LMs: Transformer-based LM with 400M parameters (TRANS-L), Transformer-based LM with 55M parameters (TRANS-S), and an LSTM-based LM with 55M parameters (LSTM) (Vaswani et al., 2017; Hochreiter and Schmidhuber, 1997). They were trained with about 3M paragraphs from Japanese newspapers and Wikipedia (3.4GB before any tokenization) with 100K parameter updates. The input was segmented into morphemes by JUMAN (Kawahara and Kurohashi, 2006), and further into subwords by sentencepiece (Kudo and Richardson, 2018), where the

[9] We tentatively adopted the **difference** of the ratio rather than, for example, ratio $\Delta_{\mathrm{human}} = r^{\mathrm{C+S+A}}_{\mathrm{human}} / r^{\mathrm{S+A}}_{\mathrm{human}}$. At least in the case of ratio, it is counter-intuitive to assume a change of 2 when the votes for TOP increase from 1 to 2, and 1.2 when they increase from 5 to 6.

Model	Setting	ρ_r	ρ_Δ	Macro F1	TOP F1	NOM F1
TRANS-L	C+S+A	0.67	−0.12	83.5	88.8	78.3
	S+A	0.60		81.7	87.6	75.8
TRANS-S	C+S+A	0.72	−0.07	85.3	89.5	81.1
	S+A	0.61		83.7	88.1	79.3
LSTM	C+S+A	0.69	−0.20	81.9	86.9	77.0
	S+A	0.62		82.3	87.1	77.5
Human	C+S+A	-	-	(100)	(100)	(100)
	S+A	-		81.1	86.5	75.7

Table 5: Results for the challenging set. The ρ_r denotes the rank correlation coefficient of the topicalization ratio exhibited by humans and LMs in each setting. The ρ_Δ denotes the rank correlation coefficient of the change in the topicalization ratio due to the presence of inter-sentential context in humans and LMs. The F1 scores were calculated with the topicalization judgment in the original text.

unigram model was used (Kudo, 2018).[10] Their hyperparameters are provided in Appendix.

Preferences of LMs. Analogous to Equation 1, we quantified the preferences of topicalization in LMs. In the S+A setting, for example, we created the MDTP of $s_{\texttt{TOP}}$ and $s_{\texttt{NOM}}$ for each instance $(c, s, a)_d$. Then, aligned with Equation 1, we calculated **topicalization ratio** of LMs as follows:

$$r_{\mathrm{LM}} = \frac{n(s_{\texttt{TOP}})}{n(s_{\texttt{TOP}}) + n(s_{\texttt{NOM}})} \ ,$$
$$n(s) := \prod_{i=1}^{|s|} p(w_i|w_{<i})^{\frac{1}{|s|}} \ , \qquad (3)$$

where $n(s)$ is the generation probability of a given sentence computed by an LM.

For each data point, two variants of topicalization ratio r_{LM} were calculated: $r_{\mathrm{LM}}^{\mathrm{C+S+A}}$ with the context c, and $r_{\mathrm{LM}}^{\mathrm{S+A}}$ without c. The score $r_{\mathrm{LM}}^{\mathrm{S+A}}$ was calculated with Equation 3, while $r_{\mathrm{LM}}^{\mathrm{C+S+A}}$ was calculated using conditional probabilities $n(s|c) = \prod_{i=1}^{|s|} p(w_i|c, w_{<i})^{\frac{1}{|s|}}$, instead of $n(s)$, in Equation 3.

Metrics We tested the LMs in terms of whether the level of certainty in choosing TOP was human-like. Specifically, Spearman's rank correlation coefficient between $r_{\mathrm{human}}^{\mathrm{C+S+A}}$ and $r_{\mathrm{LM}}^{\mathrm{C+S+A}}$ (henceforth, ρ_r) was measured. We expect that the more humans prefered TOP the more did LMs too.

Furthermore, to analyze whether the sensitivity of LMs to the context is human-like, we computed the change in LMs' topicalization preferences, analogous to Equation 2:

$$\Delta_{\mathrm{LM}} = r_{\mathrm{LM}}^{\mathrm{C+S+A}} - r_{\mathrm{LM}}^{\mathrm{S+A}}.$$

To quantify the similarity of context-sensitive preference change in LMs and humans, the rank correlation coefficients between Δ_{human} and Δ_{LM} were reported (henceforth, ρ_Δ).

Additionally, we reported F1 scores of the humans and LMs, regarding the particle choices (i.e., TOP or NOM) in the original text as the gold reference. Here, humans/LMs were considered to choose TOP when the topicalization ratio $r_{\mathrm{human/LM}}$ exceeded 0.5; otherwise, they were considered to choose NOM. Note that the corpus is from newspapers; more or less, these judgments reflect the proper use of the Japanese topic marker.

5.2 Results

Table 5 shows the results for the challenging set. In all the settings, we observed moderate correlations between humans and LMs with respect to the topicalization preferences (i.e., the ρ_r column).

Non-human-like context sensitivity. Despite the somewhat high correlations in topicalization preferences ρ_r, the correlation of context-sensitivity between humans and LMs ρ_Δ was negative; the changes in the topicalization preferences due to the presence of inter-sentential context differs between humans and LMs. This difference reveals that the context use of LMs and humans has substantial gaps.

[10] We used character coverage=0.9995, vocab size=100,000.

Model	Setting	ρ_r	ρ_Δ	Macro F1	TOP F1	NOM F1
TRANS-L	C+S+A	0.80	−0.04	88.1	89.3	86.8
	S+A	0.78		87.5	88.8	86.2
TRANS-S	C+S+A	0.80	−0.02	87.7	88.8	86.5
	S+A	0.78		87.3	88.3	86.3
LSTM	C+S+A	0.79	−0.01	85.2	86.4	84.1
	S+A	0.78		85.3	86.4	84.2
Human	C+S+A	-	-	89.7	91.0	88.4
	S+A	-		89.1	90.4	87.8

Table 6: Results for the whole dataset. The ρ_r denotes the rank correlation coefficient of the topicalization ratio exhibited by humans and LMs in each setting, and ρ_Δ denotes the rank correlation coefficient of the change in the topicalization ratio due to the presence of the inter-sentential context in humans and LMs. The F1 scores were calculated with the topicalization judgment in the original text.

LMs' insensitivity to inter-sentential context. Focusing on the differences in the F1 scores between C+S+A and S+A settings, little difference exists among the LM results. In addition, LMs typically outperform humans in terms of the F1 scores in the S+A setting. Specifically, as for the LSTM results, we obtained a somewhat strange tendency that better topicalization decisions were made better when the context is not considered. These results raise the suspicion that LMs might have made the topicalization decision with some non-contextual cues and might have performed generalizations different from those of humans.

Examples. Table 1 shows several examples of preferences of humans and LMs toward topicalization. The third example in Table 1, which belongs to the challenging set, highlighted the discrepancy between humans and LMs with respect to the context-(in)dependent preferences for topicalization. While humans chose TOP only when considering the preceding context, preferences of LMs did not change when the context information was provided.

Results in the whole dataset. We also benchmarked the topicalization judgments of the LMs in the whole dataset that we created by crowdsourcing, without limiting to the challenging set. Table 6 shows the results. Notably, even in the whole dataset, there is almost no correlation of context-sensitivity ρ_Δ between humans and LMs.

Model	ρ_r	F1		
		Macro	TOP	NOM
TRANS-L	0.18	63.8	71.1	56.5
TRANS-S	0.18	65.9	72.3	59.4
LSTM	0.17	65.1	71.9	58.4
Human	-	36.8	14.0	59.3

Table 7: The results for the A setting, where the nominative constituent alone was shown to the subjects for making the topicalization judgment.

6 Analysis

Our experiments showed that the topicalization decision of LMs is unreasonably decontextualized. To further understand such non-human-like behaviors of LMs, we investigated their behavior in situations with extremely limited context information.

As introduced in Section 3.4, we observed that humans struggle with topicalization judgment in the setting where the nominative constituent alone is shown (the A setting; Table 2). We tested whether LMs also exhibit such human-like difficulties when dealing with the whole dataset under the A setting. Notably, the purpose of this analysis is to find the situation in which LMs deviate from humans.

Language model preference. For each nominative argument a in the dataset, we computed the topicalization ratio when only the a is shown to LMs. Specifically, LMs computed the generation probability of, for example, *Kabin-ga* and *Kabin-wa*, regardless of any context, and subsequently, we compared these probabilities as per Equation. 3.

Results. Table 7 shows the results of the topicalization preferences when only the nominative constituent was shown to LMs/humans. First, the F1 scores of the LMs were surprisingly higher than those of the humans. This suggests that LMs could somehow predict the postpositional particle (TOP or NOM) without access to contextual information, although topicalization is a discourse-level phenomenon. Second, the correlation of topicalization ratio between LMs and humans was quite low. This result suggests that topicalization preferences of the LMs for nominative constituents deviate from those of humans.

7 Discussion

Contributions to linguistics. We posit that our dataset itself is also valuable for linguistic studies. Topicalization in Japanese has captured the attention in the field of linguistics for more than half a century (Matsushita, 1930; Kuno, 1973; Noda, 1996), but resources created using a large-scale corpus and crowdsourcing have been limited.

The observed relationship between mention frequency and topicalization preference (Table 4), for example, could be viewed as empirical support for the relationship between the newness of information and topicalization preference (Matsushita, 1930). The whole dataset also suggested that the intra-sentential context provides informative clues for topicalization judgment, regarding the relatively high agreement in the S+A setting. This implies that the use of TOP could not often be aligned to context-dependent topicalization phenomena; this view is consistent with Imamura et al. (2014). Note that such context-independent instances are excluded from the challenging set.

Testing coherence models. While we evaluated only the vanilla LMs, there are several options for coherence modeling, such as entity-based coherence models (Barzilay and Lapata, 2008) and neural LMs that are explicitly trained with coherence objectives (Jwalapuram et al., 2022). It would be interesting to investigate whether such models exhibit more human-like behaviors in terms of topicalization preference.

8 Conclusions

We have compared the preferences of LMs and humans for topicalization, an essential aspect of discourse. The results suggest that there exists a discrepancy between humans and LMs with respect to the generalizations for topicalization. This implication leads us to future research: *what type of inductive biases in model architecture or training objectives can lead to more human-like generalizations in discourse-level linguistic aspects?*

Acknowledgments

This work was supported by Grant-in-Aid for JSPS Fellows Grant Number JP20J22697, and JST CREST Grant Number JPMJCR20D2.

Ethical considerations

We released a new annotation layer (topicalization preference) for the existing Japanese corpus (NTC). Such preference data do not introduce any harmful content nor privacy information associated with crowd workers to the original corpus. The source NTC corpus was created using a certain Japanese newspaper (Mainichi Shinbun), which has been widely used in the NLP field (e.g., Japanese predicate argument structure analysis).

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *Proceedings of ICLR Workshop*. OpenReview.net.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34:1–34.

Wallace L. Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of ACL (System Demonstrations)*.

T Givón. 1983. *Topic Continuity in Discourse: A quantitative cross-language study*. John Benjamins.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Eva Hajicová and Jiří Mírovskỳ. 2018. Discourse coherence through the lens of an annotated text corpus: A case study. In *Proceedings of LREC*, pages 1637–1642.

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

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, pages 4129–4138.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of ACL*, pages 1725–1744.

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

Satoshi Imamura, Yohei Sato, and Masatoshi Koizumi. 2014. Influence of Information Structure on Word Order Change and Topic Marker WA in Japanese. In *Proceedings of PACLIC*.

László A. Jeni, Jeffrey F. Cohn, and Fernando De la Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *In Proceedings of ACII*, pages 245–251.

Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin. 2022. Rethinking self-supervision objectives for generalizable coherence modeling. In *Proceedings of ACL*, pages 6044–6059.

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

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of NAACL-HLT*, pages 3849–3864.

Klaus Krippendorff. 2004. *Content Analysis: an Introduction to its Methodology*. Sage publications.

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

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

Susumu Kuno. 1973. *Nihon bunpō kenkyū (Study of Japanese Grammar)*. Taishukan Shoten.

Murathan Kurfalı and Robert Östling. 2021. Probing multilingual language models for discourse. In *Proceedings of RepL4NLP*, pages 8–19.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

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

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of EMNLP*, pages 6862–6868.

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

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

Daizaburo Matsushita. 1930. *Hyōjun nihon kougohou (Standard Japanese Spoken Method)*. Chūbunkan Shoten.

Eleni Miltsakaki. 1999. Locating topics in text processing. In *CLIN*, pages 127–138.

Hisashi Noda. 1996. *Wa to ga (Wa and ga)*. Kurosio Publishers.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of CoNLL*, pages 367–379.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020. Pareto probing: Trading off accuracy for complexity. In *Proceedings of EMNLP*, pages 3138–3153.

David E Rumelhart and James L McClelland. 1985. On learning the past tenses of english verbs. Technical report, California Univ., San Diego, La Jolla. Inst. for Cognitive Science.

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

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. Probing for referential information in language models. In *Proceedings of ACL*, pages 4177–4189.

Kazuhiro Teruya. 2004. Metafunctional profile of the grammar of japanese. *Language typology. A functional perspective*, pages 185–254.

Kazuhiro Teruya. 2007. *A systemic functional grammar of Japanese*. Bloomsbury Academic.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of EMNLP*, pages 977–982.

Enric Vallduví. 1990. *The Informational Component*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

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

Marilyn Walker, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–231.

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

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

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of AAAI*, volume 34, pages 9733–9740.

Setting	#labels	class distribution		
		TOP	Both okay	NOM
C+S+A	1,853	65.7%	1.73%	32.5%
S+A	1,900	54.9%	8.47%	36.6%
A	1,732	2.02%	97.1%	0.88%

Table 8: Statistics for the challenging set. The #labels denote the number of workers who annotated the labels remaining after the post-processing. The class distribution columns denote the percentages of TOP, "Both okay", and NOM in the answers.

Parameters		TRANS-L	TRANS-S	LSTM
Fairseq model	architecture	transformer_lm_gpt2_small	transformer_lm_gpt	lstm_lm
	adaptive softmax cut off	50,000, 140,000	50,000, 140,000	50,000, 140,000
	share-decoder-input-output-embed	True	True	True
	embed_dim	1,024	384	400
	ffn_embed_dim	4,096	2048	-
	hidden_size	-	-	1,024
	layers	24	8	2
	heads	16	6	-
	dropout	0.1	0.1	0.1
	attention_dropout	0.1	0.1	-
Optimizer	algorithm	AdamW	AdamW	AdamW
	learning rates	5e-4	5e-4	1e-3
	betas	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)
	weight decay	0.01	0.01	0.01
	clip norm	0.0	0.0	0.0
Learning rate scheduler	type	inverse_sqrt	inverse_sqrt	inverse_sqrt
	warmup updates	4,000	4,000	4,000
	warmup init lrarning rate	1e-7	1e-7	1e-7
Training	batch size	61,440 tokens	61,440 tokens	20,480 tokens
	sample-break-mode	none	none	none

Table 9: Hyperparameters of the language models.