Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference

Hitomi Yanaka[1] **and Koji Mineshima**[2]
[1]The University of Tokyo, [2]Keio University
hyanaka@is.s.u-tokyo.ac.jp, minesima@abelard.flet.keio.ac.jp

Abstract

Despite the success of multilingual pre-trained language models, it remains unclear to what extent these models have human-like generalization capacity across languages. The aim of this study is to investigate the out-of-distribution generalization of pre-trained language models through Natural Language Inference (NLI) in Japanese, the typological properties of which are different from those of English. We introduce a synthetically generated Japanese NLI dataset, called the Japanese Adversarial NLI (JaNLI) dataset, which is inspired by the English HANS dataset and is designed to require understanding of Japanese linguistic phenomena and illuminate the vulnerabilities of models. Through a series of experiments to evaluate the generalization performance of both Japanese and multilingual BERT models, we demonstrate that there is much room to improve current models trained on Japanese NLI tasks. Furthermore, a comparison of human performance and model performance on the different types of garden-path sentences in the JaNLI dataset shows that structural phenomena that ease interpretation of garden-path sentences for human readers do not help models in the same way, highlighting a difference between human readers and the models.

1 Introduction

Generalization is one of the essential components that account for the understanding of language. In recent years, pre-trained models such as BERT (Devlin et al., 2019) have provided high performance on both English benchmarks (Wang et al., 2019) and multilingual benchmarks (Liang et al., 2020), suggesting that they might have some cross-lingual generalization capacity. Yet, while these models have achieved remarkable performance on an in-distribution test set (i.e., training and test splits are given as the same distribution), several previous studies have pointed out that the models fail on out-of-distribution test sets (i.e., examples drawn from a distribution different from that of the training set) (Marvin and Linzen, 2018; McCoy et al., 2019) and that the models varied widely in terms of the generalization performance (McCoy et al., 2020; Yanaka et al., 2020). It remains an open question to what extent pre-trained models can realize human-like generalization ability.

A standard task for assessing whether pre-trained language models possess human-like language understanding is Natural Language Inference (NLI), which is the task of judging whether a premise sentence entails a hypothesis sentence. Recently, a number of studies have sought to probe the generalization performance of models and detected their fallible heuristics with various NLI datasets (Naik et al., 2018; Glockner et al., 2018; McCoy et al., 2019; Rozen et al., 2019; Goodwin et al., 2020; Yanaka et al., 2021). However, these studies tend to focus on English NLI datasets, and independent analysis in multiple languages would be desirable. In response to this challenge, the study of the out-of-distribution generalization ability of NLI models from cross-lingual perspectives has begun to be explored (Hu et al., 2021) but is not yet fully developed.

The aim of this paper is to investigate to what extent pre-trained language models have generalization capacity in Japanese NLI. For this purpose, we present a Japanese linguistically challenging NLI dataset, called the Japanese Adversarial NLI (JaNLI) dataset.[1] This dataset is inspired by the English HANS dataset (McCoy et al., 2019) and is designed to cover a variety of linguistic phenomena specific to Japanese, a language typologically different from English (Hinds, 1986; Shibatani, 1990). Generating inference examples in a controlled way allows us to analyze whether language models are sensitive to factors such as word order and syntactic

[1]The dataset will be publicly available at https://github.com/verypluming/JaNLI.

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 337–349

structure in Japanese.

We present a series of experiments with the JaNLI dataset to evaluate the generalization performance of Japanese and multilingual BERT models. In addition, we compare human performance with model performance. Our experiments shed light on several shortcomings of the models and highlight the following challenges for cross-lingual generalization in NLI:

- Japanese and multilingual NLI models trained on Japanese NLI datasets other than JaNLI behave differently across different classes of sentences. In particular, cross-dataset generalization on non-entailing pairs is weaker on non-entailing pairs than on entailing pairs (Section 4.2).
- Data augmentation with a small subset of the JaNLI dataset can help improve model performance, but the accuracy is not increased for some linguistic phenomena (Section 4.3).
- Whereas humans can achieve near-perfect performance on the JaNLI dataset, there is substantial room for improving the models for Japanese NLI. In addition, structural phenomena that ease the interpretation of garden-path sentences for human readers do not help the models in the same way (Section 4.4).

2 Related Work

Previous studies have probed pre-trained language models on various NLI tasks and discovered that generalization capacity is limited for understanding diverse linguistic phenomena (Naik et al., 2018; Glockner et al., 2018; McCoy et al., 2019; Rozen et al., 2019; Goodwin et al., 2020; Yanaka et al., 2020; Hu et al., 2021; Yanaka et al., 2021) and annotation artifacts (Gururangan et al., 2018) in standard English NLI datasets such as the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets. The work most closely related to ours is HANS (McCoy et al., 2019), which is an NLI dataset designed to analyze whether models use structural heuristics to make predictions. Recently, HANS has been used for out-of-distribution evaluation data (Utama et al., 2020; Tu et al., 2020; Yaghoobzadeh et al., 2021; Du et al., 2021) and has been used for data augmentation to improve the generalization performance of models (Min et al., 2020).

Although the generalization capacity of NLI models has been studied mainly in English, non-English NLI datasets (Ham et al., 2020; Hu et al., 2020; Wijnholds and Moortgat, 2021) and multilingual NLI datasets (Conneau et al., 2018) have recently been developed to analyze the performance of pre-trained language models across languages. Several Japanese NLI datasets have been created. The Japanese SNLI dataset (Yoshikoshi et al., 2020) was generated by automatic translation of the English SNLI dataset. The Japanese Realistic Textual Entailment Corpus (Hayashibe, 2020) was created by using realistic texts (hotel reviews) and annotating labels via crowdsourcing. JSeM (Kawazoe et al., 2017) is a Japanese version of the FraCaS test suite (Cooper et al., 1994), which contains manually designed problems involving semantic phenomena that have been well studied in formal semantics. Our dataset is designed to assess whether models capture linguistic structures in Japanese or simply rely on fallible heuristics.

Recent work (Sinha et al., 2021a,b; Gupta et al., 2021; Pham et al., 2021) has shown that shuffled word order has little effect during training or inference with pre-trained language models, which in turn indicates that the models are insensitive to word order in NLI tasks. In English, however, the shuffled data are usually unacceptable and tend to obscure their gold labels. Kuribayashi et al. (2021) have re-analyzed the hypothesis that language models with lower perplexity are more human-like language models in Japanese rather than in English, and their experiments have demonstrated the lack of universality of this hypothesis and the importance of cross-lingual evaluation of models. Japanese word order is fairly free, which enables us to produce grammatically correct sentences even when the word order is shuffled. Analyzing the behavior of models on controlled Japanese inference examples should provide further insights into the sensitivity of the models to word order.

3 Dataset Generation

To analyze the generalization capacity of NLI models, we introduce a synthetically generated Japanese NLI dataset where each pair (P, H) of a premise and hypothesis is tagged with a label for *structural pattern* and *linguistic phenomenon*. Table 1 shows the definition of each pattern and some examples.

Pattern/Description	Example
FULL OVERLAP *P* and *H* share all words and differ only in word order.	Phenomena: Scrambling (Particle-swapping) *P*: ライダー が サーファー を 助け出した **rider ga surfer o rescued** (*The rider rescued the surfer*) *H*: ライダー を サーファー が 助け出した **rider o surfer ga rescued** (*The surfer rescued the rider*)
ORDER-PRESERVING SUBSET All the words in *H* are contained in *P* in an order-preserving way.	Phenomena: NP-coordination (disjunction) *P*: 学生 か 子供 が 遊んでいる **student** or child **ga playing** (*The student or the child is playing*) *H*: 学生 が 遊んでいる **student ga playing** (*The student is playing*)
MIXED SUBSET All the words in *H* are contained in *P* in a mixed (non-order-preserving) way.	Phenomena: Garden-path *P*: 子供 が 泳いでいる 学生 を 助け出した **child ga** swimming **student o rescued** (*The child rescued the swimming student*) *H*: 子供 を 学生 が 助け出した **child o student ga rescued** (*The student rescued the child.*)
SUBSEQUENCE *H* is a contiguous subsequence of *P* but not a constituent of *P*.	Phenomena: Garden-path *P*: 男の子 が 眠っている 女の子 を 見ている **boy ga sleeping** girl o looking (*The boy is looking at the sleeping girl*) *H*: 男の子 が 眠っている **boy ga sleeping** (*The boy is sleeping*)
CONSTITUENT *H* is a constituent of *P*.	Phenomena: Modal *P*: ひょっとしたら 子供 が 眠っている perhaps **child ga sleeping** (*Perhaps the child is sleeping*) *H*: 子供 が 眠っている **child ga sleeping** (*The child is sleeping*)

Table 1: Five patterns of structural relations between premise (*P*) and hypothesis (*H*) sentences: All the examples are *non-entailment*. が(ga) is a nominative case marker; を(o) is an accusative case marker.

3.1 Structural patterns and heuristics

We classify the structural relationship between premise and hypothesis sentences into five patterns, each of which is associated with a type of *heuristic* that can cause incorrect prediction of the entailment relation. For instance, a model that relies on the heuristics of judging an inference as *entailment* when the premise and hypothesis sentences share all the words will make an incorrect prediction for a non-entailment relation. We follow McCoy et al. (2019) for the definitions of the SUBSEQUENCE and CONSTITUENT patterns. McCoy et al. (2019) also proposed the *overlap* heuristics (*H* is constructed from words in *P*), which we divide into three types: FULL-OVERLAP, ORDER-PRESERVING SUBSET (ORDER-SUBSET in short), and MIXED-SUBSET. Note that these five patterns are defined to be mutually exclusive. This fine-grained classification of overlap is suitable for analysis taking into account the characteristics of Japanese that word order is relatively free (Hinds, 1986; Shibatani, 1990). We explore whether language models can perform better on some patterns compared with others.

3.2 Linguistic phenomena

To generate these five patterns of adversarial inferences in a controlled way, we focus on 11 categories of Japanese linguistic phenomena and constructions: garden-path sentences with noun-modifying clauses, scrambling (including particle-swapping), passive, causative, factive adverbs, factive verbs, modal, negation, NP-coordination, sentence-subordination (those corresponding to *because*-clauses and *if*-clauses), and sentence-coordination (sentence conjunction and disjunction).

For each phenomenon, we fix a template for the premise sentence P and create multiple templates for hypothesis sentences H. In total, we produced 144 templates for (P, H) pairs. Each pair of premise and hypothesis sentences is tagged with an entailment label (*entailment* or *non-entailment*), a structural pattern, and a linguistic phenomenon label. Table 2 shows an example template for garden-path sentences with noun-modifying clauses. See Appendix A for examples of templates for each linguistic phenomenon.

To evaluate the performance of NLI models for Japanese, the garden-path construction as shown in Table 2 deserves special attention. In this example, the challenge is to detect the boundary of the noun-modifying clause in the premise sentence P as [child ga [$_{\text{NP}}$ running cat] o chased]. Here the noun *cat* (猫) is the head of the NP with the noun-modifier *running* (走っている); when this noun is processed in the entire sentence, the subject *child* (子供) must be reanalyzed out of the clause so that *running* applies to *cat*, not to *child*. Thus, P entails H_2 (*The cat is running*) but not H_1 (*The child is running*).

Some factors are known to facilitate the interpretation of garden-path sentences (Miyamoto, 2008). We analyze whether the model predicts the entailment labels more accurately when the inference example includes such factors. We categorized problems involving garden-path sentences according to four factors that make it easier for people to interpret them: double-o-constraints (GP-double-o) (Miyamoto, 2002), presence of punctuations (GP-punctuation), selectional preference (GP-selectional) (Inoue, 2006), and presence of the topic marker *wa* (GP-wa) (Inoue, 1991). We also include the construction where the subject and object NPs of a garden-path sentence are swapped so that no garden-path effect can occur (GP-scrambling). Table 3 shows an example of each construction.

In the psycholinguistics literature, it has been observed that the processing of garden-path sentences becomes more difficult when they contain more NP-arguments (Inoue, 1990). Thus, we tagged inference problems with the number of NP-arguments.

3.3 Dataset overview

The JaNLI dataset was automatically generated by instantiating each template 100 times, resulting in a total of 14,400 examples. Table 4 shows the statistics of the linguistic phenomena. We generated the same number of entailment and non-entailment examples for each phenomenon. Table 5 shows the statistics of the structural patterns (heuristics). Note that the ratio of entailment and non-entailment examples is not necessarily $1:1$ for each pattern. This is because we first generated the templates for each linguistic phenomenon and then annotated the structural pattern with the templates.

We used 158 words (nouns and verbs) in total. Nouns and simple verbs were selected from words that occur more than 20 times in the JSICK and JSNLI datasets. Compound verbs were selected from the Compound Verb Lexicon[2]. Each transitive and intransitive verb was selected so that every noun (basically, denoting a human) is a plausible argument of it.

4 Experiments and Analysis

4.1 Experimental setting

One of our aims is to investigate the differences in behavior between monolingual and multilingual pre-trained language models. For this purpose, we conducted experiments with BERT (Devlin et al., 2019), a standard pre-trained language model that is widely used for both multilingual and Japanese texts. We compared the difference in performance between the Japanese and multilingual BERT models, which were implemented by using the transformers framework[3]. In all experiments, we trained each model for 30 epochs with early stopping (patience = 3). We perform five runs and report the average and standard deviation of the accuracy of the models.

Model Japanese BERT is pre-trained on Japanese Wikipedia, and the model processes input texts with word-level tokenization based on a standard Japanese dictionary (ipadic) (Asahara and Matsumoto, 2003), followed by the WordPiece subword tokenization (Schuster and Nakajima, 2012), trained with whole-word masking enabled for the masked language model objective. For multilingual BERT, we used a multilingual-cased model pre-trained on Wikipedia in 104 languages including Japanese, which is more recommended over a multilingual-uncased model in the case of languages with non-Latin alphabets like Japanese.

Training data To see whether the size and quality of training data affect the performance, we

[2]https://db4.ninjal.ac.jp/vvlexicon/en/

[3]https://github.com/huggingface/transformers

Templates for P and H	Sentence Example	Phenomenon/Pattern
P: NP1 ga IV NP2 o TV-o	子供 が 走っている 猫 を 追いかけた child ga running cat o chased (*The child chased the running cat*)	Garden-path sentence
$\not\Rightarrow H_1$: NP1 ga IV	子供 が 走っている (*The child is running*)	SUBSEQUENCE
$\Rightarrow H_2$: NP2 ga IV	猫 が 走っている (*The cat is running*)	MIXED-SUBSET
$\Rightarrow H_3$: NP1 ga NP2 o TV-o	子供 が 猫 を 追いかけた (*The child chased the cat*)	ORDER-SUBSET
$\not\Rightarrow H_4$: NP1 o NP2 ga TV-o	子供 を 猫 が 追いかけた (*The cat chased the child*)	MIXED-SUBSET

Table 2: Example templates for premise and hypothesis sentences. The premise P is a garden-path sentence with a noun-modifying clause. "⇒" indicates *entailment* and "$\not\Rightarrow$" *non-entailment*.

Subcategory	Template	Example
GP-double-o	NP1 ga NP2 o TV-o1 NP3 o TV-o2	子供 が 猫 を 助けた 女の子 を 追いかけた child ga cat o rescued girl o chased (*The child chased the girl who rescued the cat*)
GP-punctuation	NP1 ga , IV NP2 o TV-o	子供 が , 走っている 猫 を 追いかけた child ga PUNCT running cat o chased (*The child chased the running cat*)
GP-selectional	NP-non-human ga IV-human NP2 o TV-o	リス が しゃべっている 女性 を 追いかけた squirrel ga talking woman o chased (*The squirrel chased the woman who was talking*)
GP-wa	NP1 wa IV NP2 o TV-o	子供 は 走っている 猫 を 追いかけた child wa running cat o chased (*The child chased the running cat*)
GP-scrambling	IV NP2 o NP1 ga TV-o	走っている 猫 を 子供 が 追いかけた running cat o child ga chased (*The child chased the running cat*)

Table 3: Example templates for variants of garden-path sentences in the premise sentence.

Linguistic Phenomenon	Examples (Templates)
GP-normal	1,600 (16)
GP-double-o	800 (8)
GP-punctuation	800 (8)
GP-selectional	800 (8)
GP-wa	800 (8)
GP-scrambling	1,600 (16)
Scrambling	1,600 (16)
Passive	400 (4)
Causative	400 (4)
Factive adverb	800 (8)
Factive verb	800 (8)
Modal	600 (6)
Negation	600 (6)
NP-coordination	1,200 (12)
Sentence-subordination	800 (8)
Sentence-coordination	800 (8)
Total	14,400 (144)

Table 4: Statistics of linguistic phenomena.

Pattern (Heuristics)	Entailment	Non-entailment	Total
FULL-OVERLAP	800	1,200	2,000
ORDER-SUBSET	1,600	800	2,400
MIXED-SUBSET	3,400	2,000	5,400
SUBSEQUENCE	200	2,000	2,200
CONSTITUENT	1,200	1,200	2,400
Total	7,200	7,200	14,400

Table 5: Statistics of structural patterns.

use two types of Japanese datasets as basic training datasets: JSICK and JSNLI (Yoshikoshi et al., 2020)[4]. Table 6 shows the data split and examples from the JSICK and JSNLI datasets. We compare the model performance on the in-distribution test sets (JSICK and JSNLI) with that on the out-of-distribution test set (JaNLI). While the JSICK and JSNLI datasets use three labels (*entailment*, *contradiction*, and *neutral*), the JaNLI dataset uses two labels (*entailment* and *non-entailment*), following the HANS dataset (McCoy et al., 2019). To calculate the model accuracy on each test set, we take the highest-scoring label out of *entailment*, *contradiction*, and *neutral* and then collapse *contradiction* and *neutral* into *non-entailment*.

JSICK was created by manually translating SICK (Marelli et al., 2014), an English NLI dataset that targets compositional inference, into Japanese by experts. Marelli et al. (2014) created the original SICK dataset by expanding and normalizing the sentences from Flickr image captions with manual rules to create inference examples involving such linguistically challenging phenomena as negation, disjunction, and active-passive alternation. We ex-

[4]https://nlp.ist.i.kyoto-u.ac.jp/index.php

Dataset	Training	Test	Creation Protocol	Classes	Inference Example	Label
JSICK	5.0K	4.9K	Manual Translation	3	*P*: 水難救助をして救命胴着を着ている人は一人もいない *Nobody is practicing water safety and wearing preservers* *H*: このグループの人々は水難救助を練習していて，救命胴着具を着ている *This group of people is practicing water safety and wearing preservers*	*Contradiction* (*Non-entailment*)
JSNLI	533K	3.9K	Automatic Translation	3	*P*: 自転車に乗る赤と白のジャケットの女性 *The woman in a red and white jacket riding a bicycle* *H*: 女性が自転車に乗る *The woman is riding a bicycle*	*Entailment*
JaNLI	-	14K	Templates	2	*P*: 走っている猫を子供が追いかけた *The child chased the running cat* *H*: 子供が走っている *The child is running*	*Non-entailment*

Table 6: Overview of the Japanese NLI datasets considered in this study.

Model	Finetuned on	Test-overall		Correct: *Entailment*					Correct: *Non-entailment*				
		In-dist.	JaNLI	Full.	Order.	Mixed.	Subseq.	Const.	Full.	Order.	Mixed.	Subseq.	Const.
Ja	JSICK (5K)	92.1±0.01	51.3±0.01	99.9±0.00	97.8±0.02	79.4±0.10	98.3±0.02	88.6±0.07	0.1±0.00	6.2±0.01	6.7±0.04	32.5±0.11	22.7±0.09
	+JaNLI (0.7K)	92.3±0.01	89.3±0.06	90.8±0.04	98.6±0.01	96.8±0.02	99.2±0.01	97.3±0.02	67.1±0.17	59.1±0.04	84.6±0.23	92.4±0.09	90.4±0.05
	JSNLI (533K)	94.5±0.00	50.4±0.00	98.6±0.02	99.0±0.01	97.2±0.02	97.7±0.02	99.6±0.00	6.8±0.06	4.6±0.04	2.6±0.03	1.1±0.02	0.1±0.00
	+JaNLI (0.7K)	95.5±0.00	72.3±0.01	71.7±0.03	88.4±0.03	81.4±0.07	85.0±0.16	92.5±0.05	53.4±0.07	46.6±0.10	69.2±0.16	48.5±0.03	67.9±0.25
Multi	JSICK (5K)	73.6±0.20	50.2±0.01	66.0±0.57	64.6±0.56	57.1±0.50	62.7±0.55	63.8±0.55	33.9±0.57	34.7±0.57	36.2±0.55	45.1±0.48	43.5±0.49
	+JaNLI (0.7K)	86.5±0.08	56.9±0.06	40.8±0.37	32.9±0.33	38.0±0.35	49.8±0.44	38.8±0.36	64.2±0.33	66.0±0.37	83.3±0.19	77.4±0.32	80.9±0.23
	JSNLI (533K)	94.6±0.01	49.7±0.00	99.0±0.01	99.2±0.01	97.3±0.01	98.8±0.01	99.2±0.01	2.0±0.02	1.6±0.01	0.8±0.01	1.2±0.01	0.8±0.01
	+JaNLI (0.7K)	94.8±0.01	56.3±0.09	26.4±0.46	30.4±0.53	28.0±0.49	26.7±0.46	28.4±0.49	79.4±0.36	76.9±0.40	82.4±0.30	26.7±0.46	79.0±0.36
Human		-	94.0±0.04	94.2±0.05	97.1±0.01	92.7 ±0.04	100.0±0.00	98.3±0.03	97.8±0.01	95.8±0.05	88.7±0.09	94.3±0.08	91.1±0.14

Table 7: Results on the JaNLI test set (average accuracy and standard deviation of five runs). The number in parentheses is the size of the dataset used for finetuning. The accuracy on the in-distribution test set (JSICK/JSNLI test sets) is calculated by translating the *contradiction* and *neutral* labels into *non-entailment*.

Model	Finetuned on	GP	Scramb.	Pass.	Caus.	Fac-adv.	Fac-v.	Modal	Neg.	NP-coord.	Subord.	Sent-coord.
Ja	JSICK	49.3±0.01	50.1±0.00	49.6±0.01	47.7±0.03	49.7±0.00	51.1±0.02	54.8±0.04	63.2±0.03	50.2±0.00	69.3±0.02	46.8±0.02
	+JaNLI	92.8±0.10	79.2±0.06	49.2±0.01	56.1±0.00	75.7±0.10	90.0±0.07	93.7±0.07	98.6±0.02	99.0±0.01	98.4±0.01	97.8±0.01
	JSNLI	50.2±0.01	52.3±0.02	45.9±0.04	49.7±0.01	51.5±0.01	51.2±0.01	49.6±0.00	50.2±0.01	51.4±0.00	50.0±0.00	49.7±0.00
	+JaNLI	70.1±0.06	65.3±0.03	41.2±0.06	50.5±0.01	67.9±0.08	70.2±0.09	71.7±0.19	87.4±0.06	76.6±0.17	88.8±0.11	79.2±0.18
Multi	JSICK	49.3±0.01	49.9±0.00	49.6±0.01	48.6±0.02	49.5±0.01	50.8±0.01	50.5±0.01	49.3±0.01	49.8±0.00	61.0±0.10	49.6±0.01
	+JaNLI	56.3±0.05	52.7±0.03	49.2±0.01	56.0±0.06	53.2±0.04	58.7±0.09	57.6±0.20	62.7±0.24	61.0±0.12	61.5±0.10	60.7±0.10
	JSNLI	49.8±0.00	50.1±0.00	48.1±0.01	49.9±0.00	50.3±0.00	50.3±0.00	49.6±0.01	45.5±0.04	50.5±0.00	49.9±0.00	50.2±0.00
	+JaNLI	54.1±0.07	53.8±0.07	48.9±0.02	50.7±0.01	52.7±0.05	53.3±0.06	55.3±0.09	62.6±0.22	54.4±0.08	54.4±0.08	54.8±0.08
Human		94.2±0.05	93.3±0.03	91.7±0.08	85.0±0.17	95.8±0.05	95.0±0.02	95.6±0.08	94.4±0.05	93.9±0.03	96.7±0.04	92.5±0.09

Table 8: Results on the JaNLI test set for each linguistic phenomenon.

plore whether training on data containing these diverse linguistic phenomena improves the performance of the models. The gold labels of inference examples in JSICK were annotated via crowdsourcing. Given that the gold label of an inference example can be changed as a result of translation due to structural and lexical differences between English and Japanese, the gold labels of JSICK were re-annotated via crowdsourcing.

JSNLI was created by automatically translating the large crowdsourced English NLI dataset SNLI (Bowman et al., 2015), into Japanese. A premise sentence in the original SNLI dataset was sourced from Flickr image captions, and workers were asked to generate a corresponding hypothesis sentence for each of the three labels. Note that the size of the SNLI training set (533K) is around 100 times larger than that of the SICK training set (50K). Thus, we consider whether the quantity of the training data improves the model performance on JaNLI. The gold labels of JSNLI are the same as those of English SNLI.

We hypothesize that even if the models trained on the basic training datasets do not perform well on the JaNLI dataset, data augmentation with a small number of JaNLI examples could help the models to learn how to solve the inferences with diverse linguistic phenomena in the JaNLI dataset. To test this hypothesis, for each baseline setting, we added a small amount of JaNLI data (700 examples)[5] during the finetuning of models and checked whether the performance of the models would be improved.

4.2 Baseline results

Table 7 shows the results for the in-distribution (JSICK and JSNLI) and out-of-distribution (JaNLI) test sets. For the five heuristics, the results on

[5]There is no overlap between the subset of JaNLI and the test set in terms of (*P*, *H*) pairs. Only five premise sentences are overlapped between the added subset and the test set, for which the labels are not biased: two of them are *entailment* and three of them are *non-entailment*.

Model	Finetuned on	Scramb.	Pass.	Caus.	Fac-adv.	Scramb.	Pass.	Caus.	Fac-adv.
		Correct: *Entailment*				Correct: *Non-entailment*			
Ja	JSICK	100.0±0.00	94.7±0.05	76.7±0.07	99.4±0.01	0.2±0.00	4.5±0.04	18.7±0.05	0.0±0.00
	+JaNLI	91.5±0.03	94.7±0.03	79.3±0.14	94.8±0.06	66.9±0.12	3.8±0.03	32.8±0.14	56.7±0.25
	JSNLI	98.1±0.03	87.2±0.15	99.0±0.01	99.3±0.01	6.5±0.07	4.7±0.07	0.3±0.00	3.7±0.03
	+JaNLI	71.3±0.03	53.8±0.34	93.0±0.06	94.9±0.02	59.1±0.12	32.0±0.18	7.5±0.08	43,5±0.19
Multi	JSICK	65.9±0.57	65.7±0.57	60.2±0.53	65.2±0.56	33.9±0.57	33.5±0.58	37.0±0.55	33.9±0.57
	+JaNLI	41.8±0.40	47.5±0.44	49.0±0.45	43.9±0.38	63.6±0.34	50.8±0.45	63.0±0.38	62.5±0.34
	JSNLI	98.5±0.02	95.8±0.02	99.8±0.01	99.1±0.01	1.6±0.01	0.3±0.00	0.0±0.00	1.6±0.01
	+JaNLI	71.3±0.03	53.8±0.34	93.0±0.06	94.9±0.02	59.1±0.12	32.0±0.18	7.5±0.08	43.5±0.19

Table 9: Details of performance for problems involving linguistic phenomena for which the model performance was not improved very much by data augmentation.

Model	Train	Correct: *Entailment*						Correct: *Non-entailment*					
		Normal	Double-o	Punct.	Select.	Wa	Scramb.	Normal	Double-o	Punct.	Select.	Wa	Scramb.
Ja	JSICK	90.2±0.09	90.8±0.10	86.8±0.11	82.9±0.15	84.1±0.13	90.6±0.08	9.3±0.07	11.9±0.11	10.2±0.08	14.1±0.13	13.8±0.11	7.2±0.06
	+JaNLI	99.0±0.00	99.2±0.01	99.4±0.01	98.8±0.01	98.6±0.02	98.7±0.01	91.2±0.13	78.3±0.32	83.0±0.27	87.8±0.19	87.8±0.19	86.9±0.14
	JSNLI	98.3±0.01	95.3±0.03	99.4±0.00	98.8±0.02	99.3±0.00	98.6±0.02	2.0±0.03	3.7±0.04	1.8±0.02	0.6±0.01	2.8±0.03	1.5±0.02
	+JaNLI	83.2±0.07	88.2±0.01	86.5±0.08	92.8±0.09	88.8±0.09	82.8±0.07	58.0±0.16	54.8±0.14	53.1±0.20	49.4±0.19	47.7±0.17	55.9±0.09
Multi	JSICK	62.7±0.55	64.0±0.56	59.8±0.53	62.9±0.55	62.4±0.54	62.5±0.55	35.2±0.56	34.2±0.57	35.8±0.56	35.8±0.56	36.2±0.55	37.9±0.54
	+JaNLI	33.8±0.35	34.8±0.39	30.8±0.28	35.4±0.33	32.4±0.33	27.8±0.32	81.2±0.26	74.9±0.36	84.0±0.19	78.7±0.26	82.8±0.20	80.6±0.24
	JSNLI	98.7±0.01	97.1±0.01	99.6±0.01	99.8±0.00	99.2±0.01	98.7±0.02	0.6±0.01	1.8±0.02	0.2±0.00	0.2±0.00	1.1±0.01	0.8±0.01
	+JaNLI	28.3±0.49	29.8±0.52	30.8±0.53	32.2±0.56	30.9±0.54	29.3±0.51	79.8±0.35	79.2±0.36	78.7±0.37	74.2±0.45	77.9±0.38	78.3±0.38
Human		95.0±0.02	96.7±0.06	100.0±0.00	98.3±0.03	98.3±0.03	97.5±0.03	90.8±0.14	96.7±0.12	91.7±0.10	91.0±0.05	95.0±0.22	96.7±0.04

Table 10: Results on the JaNLI dataset for garden-path effects.

Model	Train	Correct: *Entailment*			Correct: *Non-entailment*		
		2	3	4	2	3	4
Ja	JSICK	85.8±0.12	89.4±0.10	95.5±0.04	11.8±0.09	11.0±0.11	5.1±0.05
	+JaNLI	98.9±0.01	98.8±0.01	99.1±0.00	85.2±0.22	88.5±0.14	89.6±0.13
	JSNLI	99.2±0.00	97.1±0.03	96.7±0.02	1.9±0.02	1.5±0.02	2.7±0.03
	+JaNLI	86.3±0.08	85.6±0.04	85.5±0.04	53.0±0.17	55.8±0.08	56.1±0.10
Multi	JSICK	61.5±0.54	63.6±0.55	64.8±0.56	36.8±0.55	35.2±0.56	34.3±0.57
	+JaNLI	33.0±0.33	33.9±0.37	27.4±0.31	80.8±0.23	78.7±0.30	81.3±0.26
	JSNLI	99.5±0.01	97.3±0.02	98.0±0.02	0.4±0.00	1.2±0.02	1.5±0.02
	+JaNLI	30.5±0.53	29.2±0.51	28.4±0.49	78.4±0.37	79.1±0.36	77.2±0.39
	Human	97.7±0.01	96.7±0.10	96.7±0.00	90.0±0.01	94.4±0.05	91.1±0.08

Table 11: Results on problems involving garden-path sentences for different numbers of NPs.

JaNLI are shown for correct *entailment* and *non-entailment* labels. All the models except the multilingual BERT model trained on JSICK achieved high accuracy on their in-distribution test set. They also achieved very high accuracy for the examples where the correct label is *entailment*. By contrast, we can see that regardless of the finetuning data type and model type, BERT performed substantially worse than chance (most accuracies were close to 0% while the chance level is 50%) for the examples where the correct label is *non-entailment*. These results are more or less consistent with the results reported for the English HANS dataset (McCoy et al., 2019), suggesting that the models are fooled by the heuristics in the case of Japanese as well. As explained in more detail in Section 4.4, we also evaluated human performance using a portion of the JaNLI test set. As shown in Table 7, human performance is near perfect for both *entailment* and *non-entailment* cases. Interestingly, the most difficult pattern for humans was the MIXED-SUBSET pattern (92.7 for *entailment* and 88.7 for *non-entailment*) and the same tendency was observed in the BERT models.

Comparing the performance of the multilingual and Japanese BERT models on the JaNLI dataset, we see that the overall performance of multilingual BERT is slightly lower than that of Japanese BERT. Also, comparing the effects of finetuning with JSICK and JSNLI, we see that the performance of the model finetuned with JSICK is slightly better than that of the model finetuned with JSNLI (JSICK: 51.3%; JSNLI: 50.4%). When a portion of JaNLI was added to the training data, the difference became much larger (see Section 4.3). These results suggest that the quality of the training dataset, in particular, the diversity of linguistic phenomena, can be more effective than the quantity of data for solving linguistically challenging inferences.

Table 8 shows detailed results on the JaNLI dataset for each linguistic phenomenon. The performance for each phenomenon is near the chance

level (50%) for all baselines. As with the results for the heuristics, the accuracy for *entailment* examples was near 100%, while the accuracy for *non-entailment* examples was close to 0%. When finetuned with JSICK, accuracy tended to be slightly higher for negation and sentence-subordination than for the other phenomena.

4.3 Data augmented with JaNLI examples

As Table 7 shows, when we add a small amount of JaNLI data (700 examples) during finetuning, the performance of the Japanese BERT model improved on *non-entailment* examples, while maintaining its performance on *entailment* examples: the overall accuracy increased from 51.3% to 89.3% on JSICK and 50.4% to 72.3% on JSNLI. On the other hand, the performance of the multilingual model decreased on *entailment* examples and thus failed to consistently improve the performance on JaNLI. This suggests that the training of multilingual BERT is more unstable in learning the Japanese NLI task when compared with Japanese BERT.

For both the Japanese and multilingual BERT models, the degree of performance improvement by the data augmentation was greater when finetuned with JSICK than when finetuned with JSNLI. There are two possible reasons for this result. One is that the effect of adding JaNLI examples is larger when the total size of the dataset is smaller. The other is that the data augmentation is more effective when the linguistic diversity of the original training set is higher. It should be noted that when a portion of the JaNLI dataset was added to the training set during finetuning, both Japanese and multilingual models improved in terms of performance on the in-distribution test set (JSICK/JSNLI). This result seems to support the finding that syntactic data augmentation helps to improve the robustness of models (Min et al., 2020). However, Table 8 also shows that even when JaNLI examples were used for finetuning, the performance of multilingual and Japanese BERT models was not improved on examples involving passive, causative, factive-adverbs, and scrambling. This indicates that some linguistic phenomena are difficult to learn using only data augmentation.

Table 9 shows details of performance for the four types of problems for which the model performance was not improved very much by data augmentation. For these problems, the accuracy on *non-entailment* examples was still worse than that on *entailment* examples, with the exception of the multilingual model finetuned with JSICK. For the problems involving factive adverbs, the model failed to distinguish between non-entailment examples where only the premise contains a non-factive adverb (e.g., *Perhaps the child is sleeping* $\not\Rightarrow$ *The child is sleeping*) and entailment examples where both the premise and hypothesis contain a non-factive adverb of the same type (e.g., *Perhaps the child chased the running cat* $\Rightarrow$ *Perhaps the child chased the cat*). This result is consistent with a previous study (Gupta et al., 2021) showing that the model predictions do not change even when there are repeated phrases in an inference pair.

4.4 Comparison with human judgements

To assess the difficulty of the JaNLI dataset, we collected human judgements on a subset of the JaNLI dataset through the Japanese crowdsourcing platform Lancers[6]. We selected five examples for each of 144 templates, that is, 720 inference problems in total. We collected three annotations per pair and paid annotators $0.10 per labeled pair. The annotators were six native Japanese speakers. The quality of the annotations was maintained by asking the annotators to fully understand the guidelines until they correctly answered all 10 test questions, 5 entailment and 5 non-entailment inference examples sampled from the basic training datasets.

As mentioned in Section 4.2, the overall accuracy of human performance was very high: 94.0% (see Table 7). Also, Table 8 shows that humans can make correct judgements across all types of problems present in the JaNLI dataset, despite the fact that some of them, in particular, garden-path sentences with noun-modifying clauses, are known to be hard to process based on reading time and other tests (Miyamoto, 2008). Compared with the accuracy for other linguistic phenomenon, the accuracy for causative and passive was relatively low (85% and 91.7%, respectively). For example, humans tend to predict the label for the following entailment pair as *non-entailment*.

P: 先生 が 男の子 を 海 で 泳がせた
teacher ga boy o ocean in swim-CAUSE
(*The teacher made the boy swim in the ocean*)

[6]https://www.lancers.jp/

H: 男の子 が 海 で 泳いでいる
boy ga ocean in swimming
(*The boy is swimming in the ocean*)

One possible reason why workers judge this entailment pair as *non-entailment* is that while the hypothesis sentence *H* can be interpreted to mean that the boy is swimming spontaneously of his own will, the premise sentence *P* involving a causative verb in Japanese can be interpreted to mean that the boy is *forced* to swim against his will (Kuroda, 1965; Tsujimura, 2013). A similar additional meaning beyond a simple truth-conditional content (i.e., *affectivity*) is involved in Japanese passive constructions as well (Kuno, 1973; Kuroda, 1979). It is beyond the scope of this work to address the issue of non-truth-conditional meanings.

Garden-path effects Table 10 shows detailed results on sentences with garden-path effects. As expected, the human annotators achieved slightly higher scores on garden-path problems involving factors that make the interpretation of garden-path sentences relatively easy (double-o, punctuation, selectional restriction, the topic marker *wa*, and scrambling) compared with normal garden-path problems. By contrast, there was no consistent tendency in this regard for the predictions of the BERT models. This might indicate that, unlike humans, the models do not distinguish problems involving normal garden-path phenomena from those involving factors that make them easier to interpret.

Number of NP arguments Table 11 shows results on problems involving garden-path sentences for different numbers of NPs. While Japanese psycholinguistic studies have shown that humans tend to struggle with processing garden-path sentences when they contain more NP-arguments (Inoue, 1990), but there seems to be no such trend in the case of NLI for both human judgements and model predictions. This result indicates that the number of NP arguments (at least up to four) does not significantly affect the correctness of entailment judgements.

5 Conclusion

We introduced the JaNLI dataset, which was designed to assess the generalization capacity of pre-trained language models on NLI in Japanese. Experiments showed that both Japanese and multilingual BERT models trained with basic Japanese NLI datasets performed very poorly on the JaNLI dataset. In addition, both Japanese and multilingual models, in particular the latter, struggled with learning some Japanese linguistic phenomena even when augmented with a portion of the JaNLI dataset. This suggests that there is still much room for improving the generalization capacity of pre-trained language models. Lastly, the comparison between human performance and model performance illustrated that whereas the models failed to correctly predict labels for non-entailment examples, human judgement was near perfect. Furthermore, factors that ease the interpretation of garden-path sentences for humans do not help model predictions. Overall, our dataset illuminates the vulnerabilities of the currently standard pre-trained language models and indicate a new challenge for cross-lingual generalization of NLI.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments and suggestions. This work was partially supported by JSPS KAKENHI Grant Number JP20K19868.

References

Masayuki Asahara and Yuji Matsumoto. 2003. ipadic version 2.7.0 User's Manual. Nara Institute of Science and Technology.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. 1994. FraCaS–a framework for computational semantics. *Deliverable*, D6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & family eat word salad: Experiments with text understanding. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 12946–12954.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430.

Yuta Hayashibe. 2020. Japanese realistic textual entailment corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6827–6834.

John Hinds. 1986. *Japanese: Descriptive Grammar*. Croom Helm.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526.

Hai Hu, He Zhou, Zuoyu Tian, Yiwen Zhang, Yina Patterson, Yanting Li, Yixin Nie, and Kyle Richardson. 2021. Investigating transfer learning in multilingual pre-trained language models through Chinese natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Masakatsu Inoue. 1990. Kouzou-teki aimai bun-no rikai-ni okeru goku-no nagasa-no kouka (phrasal length effects the comprehension of structurally ambiguous sentences). In *Proceedings of the 54th Japanese Psycholingual Association*, page 678.

Masakatsu Inoue. 1991. Bun-no tougo syori-ni okeru zyosi-wa-no kinou (the funcation of the particle wa in sentence parsing). In *Proceedings of the 33rd Japanese Association of Educational Psychology*, pages 55–56.

Masakatsu Inoue. 2006. Ambiguity resolution or retention in comprehending Japanese sentences. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 13(3):353–368.

Ai Kawazoe, Ribeka Tanaka, Koji Mineshima, and Daisuke Bekki. 2017. An inference problem set for evaluating semantic theories and semantic processing systems for Japanese. In *New Frontiers in Artificial Intelligence*, pages 58–65.

Susumu Kuno. 1973. *The Structure of the Japanese Language*. MIT Press.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217.

S.-Y. Kuroda. 1965. Causative forms in Japanese. *Foundations of Language*, 1(1):30–50.

S.-Y. Kuroda. 1979. On Japanese passives. In *Exploration in Linguistics: Papers in Honor of Kazuko Inoue*, pages 305–347. Kenkyusha, Tokyo.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352.

Edson T Miyamoto. 2002. Case markers as clause boundary inducers in Japanese. *Journal of psycholinguistic research*, 31(4):307–347.

Edson T. Miyamoto. 2008. Processing sentences in Japanese. In Shigeru Miyagawa and Mamoru Saito, editors, *The Oxford Handbook of Japanese Linguistics*, pages 217–249. Oxford University Press.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. 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*, pages 1145–1160.

Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan. 2019. Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 196–205.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge University Press.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021a. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *CoRR*, cs.CL/2104.06644. Version 1.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021b. UnNatural Language Inference. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP2021)*.

Natsuko Tsujimura. 2013. *An Introduction to Japanese Linguistics*. John Wiley & Sons.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3266–3280.

Gijs Wijnholds and Michael Moortgat. 2021. SICK-NL: A dataset for Dutch natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1474–1479.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural NLI models through veridicality. In *Proceedings of the 16th Conference of the European Chapter of the Association*

for Computational Linguistics: Main Volume, pages 920–934.

Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of natural language inference datasets using machine translation (in Japanese). In *Proceedings of the 244th Meeting of Natural Language Processing*.

A Templates

Table 12 shows examples of templates to generate premises and hypothesis sentences for each linguistic phenomenon (except garden-path sentences with noun-modifying clauses, whose example is shown in Table 2). Note that the conjugation of verbs can change for causative and passive forms in Japanese. Thus, when annotating a structural pattern (heuristics) tag, if the verb stems are the same in the premise and hypothesis sentences, we take them to be the same word.

The full list of templates and lexical items can be found at https://github.com/verypluming/JaNLI.

Templates for P and H	Example	Phenomenon/Pattern
P: NP1 ga NP2 o TV-o	子供 が 女性 を 見ている child ga woman o looking (*The child is looking at the woman*)	Scrambling
$\Rightarrow H_1$: NP2 o NP1 ga TV-o	女性 を 子供 が 見ている (*The child is looking at the woman*)	FULL-OVERLAP
$\not\Rightarrow H_2$: NP1 o NP2 ga TV-o	子供 を 女性 が 見ている (*The woman is looking at the child*)	FULL-OVERLAP
$\not\Rightarrow H_3$: NP2 ga NP1 o TV-o	女性 が 子供 を 見ている (*The woman is looking at the child*)	FULL-OVERLAP
P: NP1 ga NP2 ni TV-o passive	男の子 が 若者 に 押された boy ga young-man ni push-passive (*The boy was pushed by the young man*)	Passive
$\not\Rightarrow H_1$: NP1 ga NP2 o TV-o	男の子 が 若者 を 押した (*The boy pushed the young man*)	ORDER-SUBSET
$\Rightarrow H_2$: NP2 ga NP1 o TV-o	若者 が 男の子 を 押した (*The young man pushed the boy*)	MIXED-SUBSET
P: NP1 ga NP2 o IV causative	男の子 が カップル を 笑わせている boy ga couple o laugh-causative (*The boy is making the couple laugh*)	Causative
$\not\Rightarrow H_1$: NP1 ga IV	男の子 が 笑っている (*The boy is laughing*)	ORDER-SUBSET
$\Rightarrow H_2$: NP2 ga IV	カップル が 笑っている (*The couple is laughing*)	ORDER-SUBSET
P: Factive-adverb NP1 ga IV	もしかしたら サーファー が 泳いでいる perhaps surfer ga swimming (*Perhaps the surfer is swimming*)	Factive adverb
$\not\Rightarrow H_1$: NP1 ga IV	サーファー が 泳いでいる (*The surfer is swimming*)	CONSTITUENT
P: NP1 ga IV Factive-verb	サーファー が 眠っている ことは確実だ surfer ga sleeping certain (*It is certain that the surfer is sleeping*)	Factive verb
$\Rightarrow H_1$: NP1 ga IV	サーファー が 眠っている (*The surfer is sleeping*)	CONSTITUENT
P: NP1 ga NP-place de IV MODAL	子供 が 庭 で 泣いている かもしれない child ga garden de crying might (*The child might be crying in the garden*)	Modal
$\not\Rightarrow H_1$: NP1 ga NP-place de IV	子供 が 庭 で 泣いている (*The child is crying in the garden*)	SUBSEQUENCE
$\Rightarrow H_2$: NP-place de NP1 ga IV MODAL	庭 で 子供 が 泣いている かもしれない (*The child might be crying in the garden*)	MIXED-SUBSET
P: NP1 ga NP-place de IV NEG	子供 が 海辺 で 横たわっている わけではない child ga beach de lying negation (*The child is not lying on the beach*)	Negation
$\not\Rightarrow H_1$: NP1 ga NP-place de IV	子供 が 海辺 で 横たわっている (*The child is lying on the beach*)	SUBSEQUENCE
$\Rightarrow H_2$: NP-place de NP1 ga IV NEG	海辺 で 子供 が 横たわっている わけではない (*The child is not lying on the beach*)	MIXED-SUBSET
P: NP1 ga NP2 ka NP3 o TV-o	子供 が 女性 か 男性 を 見ている child ga woman or man o looking (*The child is looking at the woman or the man*)	NP-coordination
$\not\Rightarrow H_1$: NP1 ga NP2 o tv-o	子供 が 女性 を 見ている (*The child is looking at the woman*)	ORDER-SUBSET
$\not\Rightarrow H_2$: NP1 ga NP3 o tv-o	子供 が 男性 を 見ている (*The child is looking at the man*)	ORDER-SUBSET
P: NP1 ga IV reason NP2 ga NP3 o TV-o	カップル が 遊んでいる から 子供 が ライダー を 見ている couple ga playing because child ga rider o looking (*Because the couple is playing, the child is looking at the rider*)	Sentence-subordination
$\Rightarrow H_1$: NP1 ga IV	カップル が 遊んでいる (*The couple is playing*)	CONSTITUENT
$\Rightarrow H_2$: NP2 ga NP3 o TV-o	子供 が ライダー を 見ている (*The child is looking at the rider*)	CONSTITUENT
P: NP1 ga IV ka NP2 ga NP3 o TV-o	女の子 が 走っている か 子供 が ライダー を 追い回している girl ga running or child ga rider o chasing (*The girl is running or the child is chasing the rider*)	Sentence-coordination
$\not\Rightarrow H_1$: NP1 ga IV	女の子 が 走っている (*The girl is running*)	CONSTITUENT
$\not\Rightarrow H_2$: NP2 ga NP3 o TV-o	子供 が ライダー を 追い回している (*The child is chasing the rider*)	CONSTITUENT

Table 12: Example templates for premise and hypothesis sentences for each linguistic phenomenon. "⇒" indicates *entailment* and "⇏" *non-entailment*.