JAMP: Controlled Japanese Temporal Inference Dataset for Evaluating Generalization Capacity of Language Models

Tomoki Sugimoto[1], **Yasumasa Onoe**[2], **Hitomi Yanaka**[1]
[1]The University of Tokyo, [2]The University of Texas at Austin
{sugimoto.tomoki,hyanaka}@is.s.u-tokyo.ac.jp
yasumasa@utexas.edu

Abstract

Natural Language Inference (NLI) tasks involving temporal inference remain challenging for pre-trained language models (LMs). Although various datasets have been created for this task, they primarily focus on English and do not address the need for resources in other languages. It is unclear whether current LMs realize the generalization capacity for temporal inference across languages. In this paper, we present JAMP, a Japanese NLI benchmark focused on temporal inference. Our dataset includes a range of temporal inference patterns, which enables us to conduct fine-grained analysis. To begin the data annotation process, we create diverse inference templates based on the formal semantics test suites. We then automatically generate diverse NLI examples by using the Japanese case frame dictionary and well-designed templates while controlling the distribution of inference patterns and gold labels. We evaluate the generalization capacities of monolingual/multilingual LMs by splitting our dataset based on tense fragments (i.e., temporal inference patterns). Our findings demonstrate that LMs struggle with specific linguistic phenomena, such as habituality, indicating that there is potential for the development of more effective NLI models across languages.

1 Introduction

Natural Language Inference (NLI) is the task of determining whether a set of premises entail a hypothesis. NLI involving temporal inference is a challenging task and remains a significant problem for pre-trained language models (LMs). One line of research has investigated the temporal inference abilities of LMs (Kober et al., 2019; Vashishtha et al., 2020; Thukral et al., 2021; Chen and Gao, 2022). However, existing datasets and analyses primarily focus on English, and more analysis and datasets are required for other languages, including Japanese. Therefore, it is still unclear to what extent current LMs can perform various types of

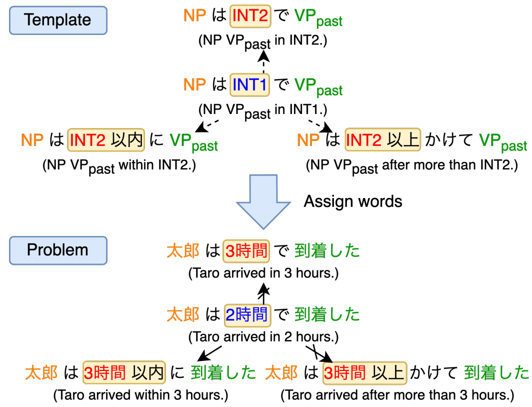


Figure 1: An illustration of our data annotation process. INT in the templates means interval. $\dashrightarrow$ means that the gold label is undetermined, $\rightarrow$ means that the gold label is *Entailment* and $\not\rightarrow$ means that the gold label is *Contradiction*.

temporal inference across languages. In this paper, we construct JAMP[1], which is a Japanese NLI dataset for temporal inference, and evaluate the generalization capacity of several LMs on our dataset.

Our goal is to construct a temporal inference dataset that precisely assesses the generalization capacities of LMs. Manual annotation is a viable option for achieving this goal, but it does not fully meet our needs based on several limitations described below. Although using crowdsourcing to increase the size of datasets may be cost-effective (Bowman et al., 2015; Williams et al., 2018), managing biases and artifacts in the resulting data can be challenging (Poliak et al., 2018b; Gururangan et al., 2018). In contrast, datasets manually constructed by experts (Cooper et al., 1996; Kawazoe et al., 2015) may have high quality but are potentially expensive to scale. Additionally, manual dataset construction makes it difficult to control the distribution of vocabulary and inference patterns in a dataset because it heavily relies on the prior knowledge of each annotator (e.g., word choice). To address the issues associated with

[1]Our dataset is available on https://github.com/tomo-ut/temporalNLI_dataset

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 57–68

Main Tense Fragment	Sub-tense Fragment		Example Problem
Temporal anaphora	Reference resolution of 昨日 (*yesterday*)	P	昨日、 APCOM は 契約書 に 署名した。 **yesterday** , APCOM wa contract ni sign . (APCOM signed the contract **yesterday**.) 今日 は 7 月 14 日 土曜日 だ。 **today** wa 7 month 14 day Saturday da . (**Today** is Saturday, July 14.)
		H	APCOM は 13 日 の 金曜日 に 契約書 に 署名した。 APCOM wa 13 day no Friday ni contract ni sign . (APCOM signed the contract on Friday the 13th.)
		G	Entailment
Interval	Completion of eventuality	P	スミス は バーミンガム に **2 年** 住んだ。 Smith wa Birmingham ni **2 year** live . (Smith lived in Birmingham **for two years**.)
		H	スミス は バーミンガム に 住んだ。 Smith wa Birmingham ni live . (Smith lived in Birmingham.)
		G	Entailment

Table 1: Examples of tense fragments and corresponding problems. P, H, and G indicate a set of premises, a hypothesis, and a gold label, respectively.

manual annotation, prior work uses template-based approaches that automatically assign diverse vocabulary to templates that are manually created by experts to construct scalable datasets (Richardson et al., 2020; Yanaka and Mineshima, 2021). By using this method, we can strictly manage the vocabulary and inference patterns in a dataset, thus it is a suitable approach for probing LMs.

Figure 1 presents our data annotation process, which consists of two stages: *template creation* and *problem generation*. We first collect Japanese temporal inference examples from JSeM (Kawazoe et al., 2015), which is the Japanese version of FraCaS (Cooper et al., 1996), and manually transform them into templates by masking content words (e.g., nouns and verbs) and temporal expressions (e.g., date and time), producing 46 tense fragments (i.e., temporal inference patterns) based on formal semantics. We then generate examples by assigning content words sampled from a Japanese case frame dictionary (Kawahara and Kurohashi, 2006) and randomly generating temporal expressions to those templates. These techniques ensure that the sentences in JAMP are diverse and cover a wide range of temporal inference patterns. It is important to note that our temporal NLI examples are derived from a diverse set of templates that are classified with tense fragments, allowing us to create different test splits depending on the goal of evaluation, such as generalization across different tense fragments.

We evaluate two Japanese models and one multilingual model on our dataset. We analyze whether they can solve our dataset in a zero-shot setting (trained on existing Japanese NLI datasets) and a fine-tuning setting (trained on a small subset of our dataset). The experimental results demonstrate that the LMs can generalize across different temporal expressions but fail to generalize some tense fragments such as habituality.

2 Background

2.1 Frame

Frame is one of the basic knowledge for language understanding. There are several English resources for frame knowledge, including VerbNet (Schuler, 2005), FrameNet (Baker et al., 1998), and PropBank (Palmer et al., 2005), and previous studies have used these resources to construct datasets (Poliak et al., 2018a; Mitra et al., 2020).

In Japanese, case particles (e.g., が–pronounced *ga*) are attached to verbal arguments (e.g., subject) and determine the case frame. A Japanese case frame dictionary (Kawahara and Kurohashi, 2006) is the largest resource that reflects these characteristics of Japanese language. This case frame dictionary is a set of 110,000 predicates and associated nouns extracted from 10 billion sentences, that are annotated for each predicate usage. Table 2 shows an example of a case frame in the Japanese case frame dictionary.

As shown in Table 2, the case frame dictionary contains information regarding the frequencies of case frames and nouns. In this paper, we use these case frames to generate a dataset containing diverse sentence patterns without grammatical errors.

2.2 Fragments

Some existing datasets (Cooper et al., 1996; McCoy et al., 2019; Yanaka and Mineshima, 2021), including JSeM (Kawazoe et al., 2015), define problem categories for each problem for further analysis. In this study, we systematically defined tense fragments (i.e., temporal inference patterns) based on

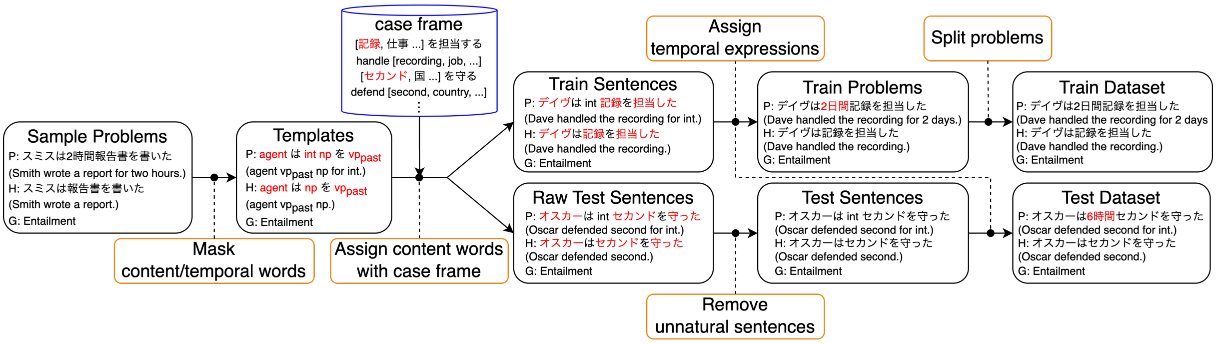


Figure 2: Overview of our data construction pipeline. 1) We first create temporal inference templates from existing examples. 2) We then assign content words using the Japanese case frame dictionary. 3) After isolating train and test examples, we assign temporal expressions to the candidate sentences. Additionally, we manually filter unusable sentences from the test examples.

到着する (arrive): verb, freq=118520	
ga	選手 (athlete)$_{freq=205}$, 大統領 (president)$_{freq=114}$, $\cdots$
ni	空港 (airport)$_{freq=24705}$, ホテル (hotel)$_{freq=9639}$, $\cdots$
$\vdots$	...
de	飛行機 (airplane)$_{freq=347}$, バス (bus)$_{freq=293}$, $\cdots$

Table 2: An example of a case frame in the Japanese case frame dictionary.

the categories of temporal inference patterns in JSeM.

Table 1 shows some examples of tense fragments (see Appendix A for additional tense fragments). In Table 1, "Main Tense Fragment" represent higher-level classifications, and "Sub-tense Fragment" represent sub-classifications that are subdivided from the main tense fragments. Tense fragments enable a more detailed analysis of LMs' understanding of temporal inference.

3 JAMP

In this paper, we present JAMP, which is a Japanese NLI dataset for temporal inference, and propose a method for automatic construction from templates based on tense fragments. Figure 2 shows the pipeline of our method. First, we create a template by masking content words and temporal expressions in existing temporal NLI problems (§3.1). A template consists of the following triplet: (i) a set of premises in which content words and temporal expressions are masked, (ii) a hypothesis in which content words and temporal expressions are masked, and (iii) a condition for determining a gold label. Here, a gold label can take on three values: *Entailment*, *Contradiction*, and *Neutral*. Next, we generate training and test sentences by assigning content words selected from the vocabulary list to the template (§3.2). We create a vocabulary list by using the Japanese case frame dictionary to make sentences more coherent.[2]

Template	P:	agent_1 が interval_1 以内に np_1 を vp_1_past。
	H:	agent_1 は interval_2 以内に np_1 を vp_1_past。
	G:	if interval_1 $\leq$ interval_2 then Entailment else Neutral
Generated Problem	P:	エレン が 6 年間 以内 に ゴール を 達成した。 Ellen ga 6 years within ni goal o achieved . (Ellen has achieved her goal within six years.)
	H:	エレン は 5 年間 以内 に ゴール を 達成した。 Ellen wa 5 years within ni goal o achieved . (Ellen has achieved her goal within five years.)
	G:	Neutral

Table 3: An example of a template and a problem generated by our method.

We manually inspect all sentences in the test examples and eliminate any sentences that are unnatural or harmful. We then generate train and test problems by assigning temporal expressions to train and test sentences. Finally, we split the training problems along three axes (e.g., tense fragment, time format, and time span) to create training data for various experimental settings (§3.4). In this section, we describe each of these steps in detail.

3.1 Template Creation

In the first step, we construct templates consisting of a set of premises, a hypothesis, and a gold label. We create templates for temporal problems based on problems in the temporal inference section of JSeM by masking content words such as nouns and verbs (e.g., スミス (*Smith*), 住んだ (*lived*)), and temporal expressions (e.g., 7 月 14 日 (*July 14*), 2 年 (2 *years*)). Additionally, because the gold label depends on the temporal expression in the sentence, we convert the original gold label into a condition in which the gold label is determined by specifying a temporal expression. Table 3 shows

[2]We considered a generation method using masked LMs or generative models but did not adopt them in this study because the generation time was too long, and it was difficult to control the vocabulary and not change inference patterns and syntactic structures.

an example of the template. In the example in Table 3, the condition is "if interval_1 $\leq$ interval_2 then *Entailment* else *Neutral*" and the gold label is determined according to temporal expressions in interval_1 and interval_2.

There can be strong correlations between specific words and labels in examples generated from templates based on certain JSeM problems. Because such correlations could introduce undesired biases into our dataset, we removed these correlations by constructing new challenging templates for some JSeM problems (see Appendix B for examples).

3.2 Problem Generation

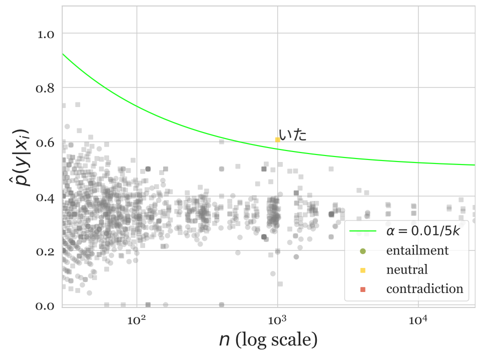
We generate problems by filling the masks in templates with various nouns, verbs, and temporal expressions and determining the gold label from these temporal expressions. We use the Japanese case frame dictionary as a vocabulary for selecting verbs and nouns (§2.1). In this study, we manually filter about 30 offensive words from verbs whose frequency in the dictionary is greater than 1000 and nouns whose frequency in the dictionary is greater than 100 extracted from the case frame dictionary and use filtered words.

We target two types of temporal expressions in this study: time points (e.g., 8 月 16 日 7 時 (*August* 16, 7:00)) and intervals (e.g., 3 ヶ月 (3 *months*)). For time points, we use 10 formats combining year/month/day/hour units: Year (Y), Month (M), Day (D), Hour (H), YM, MD, DH, YMD, MDH, and YMDH. For intervals, we use four formats: Year, Month, Day, and Hour.

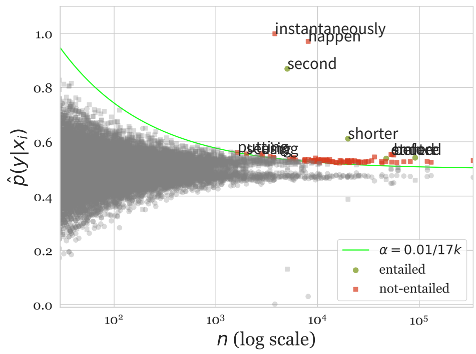
We assign content words and temporal expressions to templates as follows. First, we randomly select a verb with the case in the template from the case frame dictionary. Next, we randomly select nouns that the selected verb can take as its case in the template. Here, we select a noun for a subjective case from a manually created list of common first names (e.g., *Alice* and *Bob*).

Then, if a temporal expression exists in the original problem corresponding to the template, we generate a new temporal expression as follows and assign it to templates. If the original temporal expression is an interval, we generate an interval by concatenating an integer randomly selected from one to nine according to one of the four formats described above. If the original temporal expression is a time point, we first randomly select a time point within the range of January 1, 2000, at 0:00 to December 31, 2020, at 24:00. Then, one of the ten formats described above is applied to the selected time point. For example, if the MD format is applied to 0:00 on January 1, 2010, then the generated temporal expression will be "January 1."

Finally, we assign a gold label by evaluating the condition for the gold label in the template. Table 3 shows an example of a template and the problem generated from that template. In Table 3, the condition is "if interval_1 $\leq$ interval_2 then *Entailment* else *Neutral*." Because the generated temporal expressions for interval_1 and interval_2 are 6年間 (*six years*) and 5年間 (*five years*), respectively, its gold label is *Neutral*. To ensure that the distribution of gold labels is approximately uniform, we generate the same number of problems from each pair of a template and a gold label.



(a) JAMP



(b) Temporal NLI

Figure 3: The artifact statistics of (a) JAMP and (b) Temporal NLI (Vashishtha et al., 2020) training sets. The majority of words in JAMP, with the exception of "いた," are located below the green line, implying that they do not exhibit spurious correlations with the gold labels. A substantial number of words in Temporal NLI correlate with the gold labels.

Unnatural Sentence	Cause
チャーリー が インク を 吸った。 Charlie ga ink o sucked . (Charlie sucked ink.)	Semantically unnatural
ウォルター は 性格 に 変わった。 Walter wa characteristic ni changed . (Walter changed in character.)	Incomplete sentence
キャロル は 速度 に 生ずるていた。 Carroll wa speed ni arise . (Carroll arose to speed.)	Semantically unnatural Grammatically unnatural

Table 4: Examples of unnatural sentences we filtered.

3.3 Quality Control

3.3.1 Dataset Artifacts

Previous works have demonstrated that existing datasets are often affected by dataset artifacts and spurious correlations between surface features and gold labels (Jia and Liang, 2017; Gururangan et al., 2018; Poliak et al., 2018b). We conduct statistical analysis on our dataset following the method outlined by Gardner et al. (2021) to identify token-level artifacts. Our analysis reveals the extent to which certain words are highly correlated with one of three labels (see Appendix D for details).

Our automatic data annotation approach enables us to effectively manage the examples that we generate. We conduct this statistical analysis during the data generation phase and modify vocabulary words and templates to eliminate shortcuts and spurious correlations between certain words and gold labels. As depicted in Figure 3, the majority of words in JAMP do not exhibit spurious correlations with the gold labels, whereas a significant number of words in Temporal NLI (Vashishtha et al., 2020) correlate with the gold labels.[3] In JAMP, the word “いた”[4] stands out as an exception, but its impact is relatively low because its score is close to the green line.

3.3.2 Dataset Quality

Naturalness We manually check the naturalness of all test examples and filter out disqualified sentences (approx. 40% of all sentences).[5] Table 4 shows examples of sentences we remove from the test set and the reasons for their removal.

Semantically unnatural (e.g., the examples at the top and bottom of Table4) refers to sentences that are grammatically correct but may not be plausible. One reason for the generation of such sentences is that the Japanese case frame dictionary does not describe the correspondence between cases (e.g., ヲ格 (accusative) and ニ格 (dative)). The second case, an incomplete sentence, could be generated since the Japanese case frame dictionary does not describe the essential case for predicates. Other examples, such as the third, show verbs conjugated in the wrong form. This is probably because the verb is not included in the dictionary used to conjugate the verb.

Correctness We randomly sample 100 cases from the constructed test data and manually judge their entailment labels. We check whether the judgement is the same as their gold labels. We confirm that the gold labels in all cases were annotated as intended. However, the gold labels for some problems were debatable. For example, in the sentence *I read a book for three hours*, the meaning of *for three hours* can be interpreted as "just three hours," "about three hours," and "at least three hours". The interpretation depends on the speaker and the context. In such cases, their gold labels depend on the reading, but we confirmed that they are correct in at least one of the possible readings.

3.4 Split Problems

Our controlled data generation method enables us to split problems into seen problems (i.e., problems included in both test and training data) and unseen problems (i.e., problems included only in test data) systematically, which is suitable for investigating the generalization capacity of LMs. In this study, we split our training data to analyze whether LMs can generalize various temporal inference patterns learned from training data. We split the training data based on three axes: tense fragment, time format, and time span. Table 5, 6, and 7 show an example of a seen/unseen problem in each split.

3.4.1 Tense Fragment-Based Split

Tense fragment refers to the categorization of the problems described in Section 2.2. We define two splits based on the tense fragments: FRAGMENT_EASY and FRAGMENT_HARD. These splits aim to test whether LMs can learn temporal inference from basic problems and generalize the acquired inference patterns to more challenging problems. Therefore, both FRAGMENT_EASY and FRAGMENT_HARD include only basic problems in the training data and challenging problems in the test data. FRAGMENT_HARD contains a higher pro-

[3]We sample 100k training examples for this statistical analysis.

[4]This Japanese word has multiple grammatical roles. One is a past stative verb, and another is a past continuous form of a verb.

[5]We ask 3 graduate students studying NLP/linguistics to judge sentence quality.

	Seen problem	Unseen problem
	TF: Order relation - Transitive, Gold label: Entailment	TF: Order relation - Transitive + Before/After, Gold label: Entailment
P	マレット は イブ が 出掛ける **前** に 出掛けた。 Mallett wa Eve ga leave **before** ni leave . (Mullet left **before** Eve left.) イブ は チャーリー が 出掛ける **前** に 出掛けた。 Eve wa Charlie ga leave **before** ni left . (Eve left **before** Charlie left.)	マーヴィン は ペギー が 留学する **前** に 留学した。 Marvin wa Peggy ga study abroad **before** ni study abroad . (Marvin studied abroad **before** Peggy studied abroad.) マーヴィン は キャロル が 留学した **後** に 留学した。 Marvin wa Carol ga studied abroad **after** ni studied abroad . (Marvin studied abroad **after** Carol studied abroad.)
H	マレット は チャーリー が 出掛ける **前** に 出掛けた。 Mallett wa Charlie ga leave **before** ni leave . (Mullet left **before** Charlie left.)	ペギー は キャロル が 留学した **後** に 留学した。 Peggy wa Carol ga study abroad **after** ni study abroad . (Peggy studied abroad **after** Carol studied abroad.)
	TF: Usage of 現在 (*now*) - Present tense, Gold label: Entailment	TF: Usage of 現在 (*now*) - Past tense, Gold label: Neutral
P	マレット は 皆さん に 考え方 を 述べ**ている**。 Mallett wa everyone ni thinking o state . (Mallett **is** stating his thinking to everyone.)	アイザック は 見学 に バー を 訪れ**ていた**。 Isaac wa tour ni bar o visit . (Isaac **was** visiting the bar for a tour.)
H	マレット は **現在** 皆さん に 考え方 を 述べている。 Mallett wa **now** everyone ni thinking o state . (Mallett is **now** stating his thinking to everyone.)	アイザック は **現在** 見学 に バー を 訪れている。 Isaac wa **now** tour ni bar o visit . (Isaac is **now** visiting the bar for a tour.)

Table 5: Examples of problems that are in the training data (seen problems) and corresponding problems that are not in the training data (unseen problems) in a tense fragment-based split setting. TF means the tense fragment.

portion of challenging problems and fewer tense fragments in the training data, which is a more difficult setting for models.

We define basic and challenging problems based on the sub-tense fragments in the tense fragment classification. For example, as in the first example in Table 5, suppose a certain tense fragment has sub-tense fragments that are finer than that tense fragment. In this case, the original tense fragment (Order relation – Transitive) is considered as basic, and the subcategories (Order relation – Transitive + Before/After) are considered as challenging. In contrast, as in the second example in Table 5, if there is no such sub-tense fragment, but there are sub-tense fragments with the same granularity as that of the classification, one (Usage of 現在 (*now*) – Present tense) is considered as basic, and the other (Usage of 現在 (*now*) – Past tense) is considered as challenging.

3.4.2 Time Format-Based Split

Time format represents the format of the temporal expression inserted in a problem. In this study, we define ten time formats by combining multiple time units (year, month, day, and hour) for time points and define two splits based on the time formats. This split aims to test whether LMs can learn the size relationships between time units (year > month > day > hour) from a minimal number of combinations of units and generalize the acquired inference patterns to apply them to complex combinations.

The first split is FORMAT_HARD, which contains only a single time unit pattern (i.e., patterns involving only year, only month, only day, or only hour) in a training set and evaluates models on combined patterns of multiple time units.

The other split is FORMAT_EASY, which includes a minimum number of combinations (i.e., year-month pattern, month-day pattern, and day-hour pattern) that allow the models to understand the size relationships between time units, as shown in the second example in Table 6. By comparing the accuracy of FORMAT_EASY and FORMAT_HARD, we can determine whether LMs can learn and generalize the size relationships between time units.

3.4.3 Time Span-Based Split

Time span represents the closeness of temporal expressions when multiple temporal expressions appear in a problem. In this study, we define two time spans: SHORT and RANDOM. In SHORT time span problems, the temporal expressions are generated such that the time points included in the problem are close to each other (see Appendix C), as shown in the unseen problem in Table 7. On the other hand, in RANDOM time span problems, the distance between the time points included in the problem is not predetermined, and the temporal expressions are generated in the same manner as described in Section 3.2. Therefore, the distances between the time points included in a problem are often far apart, as shown in the seen problem in Table 7.

When a model determines the order of two time points, the model must compare the two time points in order, starting with the largest unit. If two time points are far apart, then the model can determine their order by comparing only the larger units, but if two time points are close, then the model must compare additional units to determine their order. For example, the order of January 1, 2010, at 1:00 and October 10, 2020, at 10:00 can be determined by looking only at the year, but the order of January 1, 2010, at 1:00 and January 1, 2010, at 10:00

	Seen problem	Unseen problem
	Format: Year, Gold label: Neutral	Format: Year-Month-Day-Hour, Gold label: Entailment
P	パット が 6 年間 以内 に 代価 を 支払った。 Pat ga 6 year within ni price o paid . (Pat paid the price within 6 years.) パット は **2009** 年 に その 代価 を 支払い 始めた。 Pat wa **2009 year** ni its price o pay began . (Pat began paying the price in **2009**.)	エレン が 2 年間 以内 に 考え を 変えた。 Ellen ga 2 years within ni mind o changed . (Ellen changed her mind within 2 years.) エレン は **2016** 年 **11** 月 **18** 日 **15** 時 に その 考え を 変え 始めた。 Ellen wa **2016 year 11 month 18 day 15 hour** ni its mind o change began . (Ellen began to change her mind at **15:00 on November 18, 2016**.)
H	パット は **2011** 年 まで に その 代価 を 支払い 終えた。 Pat wa **2011 year** until ni its price o pay finished . (Pat finished paying the price by **2011**.)	エレン は **2020** 年 **10** 月 **15** 日 **21** 時 まで に その 考え を 変え 終えた。 Ellen wa **2020 year 10 month 15 day 21 hour** until ni its mind wo change finished . (Ellen finished changing her mind by **21:00 on October 15, 2020**.)
	Format: Year-Month, Gold label: Entailment	Format: Year-Month-Day-Hour, Gold label: Entailment
P	**2018** 年 **8** 月 以来、 ウォルター は 閣僚 に 指示している。 **2018 year 8 month** since , Walter wa cabinet ni instruct . (Since **August 2018**, Walter has instructed cabinet members.) 現在、 **2018** 年 **11** 月 である。 now , **2018 year 11 month dearu** . (It is now **November 2018**.)	**2008** 年 **2** 月 **27** 日 **0** 時 以来、 ビクター は ソフトバンク に 移籍している。 **2008 year 2 month 27 day 0 hour** since , Victor wa Softbank ni transfer . (Since **0:00 on February 27, 2008**, Victor has been transferred to Softbank.) 現在、 **2008** 年 **2** 月 **27** 日 **4** 時 である。 now , **2008 year 2 month 27 day 4 hour dearu** . (It is now **4:00 on February 27, 2008**.)
H	ウォルター は **2018** 年 **9** 月 には 閣僚 に 指示していた。 Walter wa **2018 year 9 month** niwa cabinet ni instruct . (Walter had instructed the cabinet ministers in **September 2018**.)	ビクター は **2008** 年 **2** 月 **27** 日 **1** 時 には ソフトバンク に 移籍していた。 Victor wa **2008 year 2 month 27 day 1 hour** niwa Softbank ni transfer . (Victor was transferred to Softbank at **1:00 on February 27, 2008**.)

Table 6: Examples of problems that are in the training data (seen problems) and corresponding problems that are not in the training data (unseen problems) in a time format-based split setting.

	Seen problem	Unseen problem
	Span: Random, Gold label: Neutral	Span: Short, Gold label: Contradiction
P	**2002** 年 **8** 月 **16** 日 **7** 時 以来、 ウォルター は 実家 に 泊まっている。 **2018 year 8 month 16 day 7 hour** since , Walter wa parents' house ni stay . (Walter has been staying at his parents' house since **7:00 on August 16, 2002**.) 現在、 **2013** 年 **5** 月 **26** 日 **3** 時 である。 now , **2013 year 5 month 26 day 3 hour** dearu . (It is now **3:00 on May 26, 2013**.)	**2015** 年 **9** 月 **11** 日 **7** 時 以来、 フランク は 細工 に 挑戦している。 **2015 year 9 month 11 day 7 hour** since , Frank wa craft ni try . (Frank has been trying to craft since **7:00 on September 11, 2015**.) 現在、 **2015** 年 **9** 月 **11** 日 **10** 時 である。 now , **2015 year 9 month 11 day 10 hour** dearu . (It is now **10:00 on September 11, 2015**.)
H	ウォルター は **2018** 年 **5** 月 **15** 日 **12** 時 には 実家 に 泊まっていた。 Walter wa **2018 year 5 month 15 day 12 hour** niwa parents' house ni stay . (Walter was staying at his parents' house at **12:00 on May 15, 2018**.)	フランク は **2015** 年 **9** 月 **11** 日 **5** 時 には 細工 に 挑戦していた。 Frank wa **2015 year 9 month 11 day 5 hour** niwa craft ni try . (Frank was trying to craft at **5:00 on September 11, 2015**.)

Table 7: Examples of problems that are in the training data (seen problems) and corresponding problems that are not in the training data (unseen problems) in a time span-based split setting.

requires comparing the year, month, day, and hour in order. Therefore, we consider that determining the order relationships between close time points is more difficult than determining the order relationships between distant time points.

We define a time span-based split that contains only RANDOM in the training data. This split aims to test whether LMs can learn the order relationships of temporal expressions and generalize the acquired inference patterns to apply them to combinations of temporal expressions that require more difficult evaluation.

4 Experiments

We evaluate several NLI models on our dataset. We consider six pre-trained LMs (Japanese BERT-base/large, Japanese RoBERTa-base/large, multilingual XLM-RoBERTa-base/large)[6] available on huggingface/transformers[7] in our experiments. We conduct experiments in three settings: zero-shot (monolingual), zero-shot (cross-lingual), and fine-tuning. Here, zero-shot means that we do not use our training data but use existing Japanese NLI datasets for training data. The statistics of the datasets used in our experiments are provided in Appendix E.

Zero-shot setting (monolingual) We train the LMs on three concatenated NLI datasets: the standard Japanese NLI datasets JSNLI (automatic translation of the English SNLI dataset (Bowman et al., 2015)) (Yoshikoshi et al., 2020) and JSICK (manual translation of the English SICK dataset (Marelli et al., 2014)) (Yanaka and Mineshima, 2022), and the Japanese NLI dataset PLMUTE_ja (Sugimoto and Yanaka, 2022), which involves temporal order. We then evaluate the models on our test data.

Zero-shot setting (cross-lingual) We train the LMs on three concatenated NLI datasets: the standard English NLI dataset SNLI, SICK, and the English NLI dataset PLMUTE (Thukral et al., 2021), which involves temporal order and duration. We then evaluate the models on our test data.

Fine-tuning setting We train and evaluate the LMs on our training data and test data.

Additionally, in the fine-tuning setting, we train the LMs on the split training data described in Sec-

[6] We did not evaluate the prompt-tuning models such as GPT-3 because accurate comparisons with other models in the fine-tuning setting are difficult.

[7] https://huggingface.co/transformers/

Model		seen/ unseen	Zero-shot		Fine-tuning						
			Mono lingual	Cross-lingual	IID Split	Tense Fragment		Time Format			Time Span
						Easy	Hard	Easy	Hard	Δ	
BERT	base	seen	-	-	.891$_{\pm 0.02}$	.879$_{\pm 0.01}$	.812$_{\pm 0.05}$	.839$_{\pm 0.02}$	.800$_{\pm 0.02}$	.039$_{\pm 0.03}$	.757$_{\pm 0.03}$
		unseen	.428$_{\pm 0.02}$	-	-	.405$_{\pm 0.04}$	.379$_{\pm 0.02}$	.897$_{\pm 0.03}$	.761$_{\pm 0.04}$	**.136**$_{\pm 0.05}$	.662$_{\pm 0.05}$
		Δ	-	-	-	**.474**$_{\pm 0.04}$	**.433**$_{\pm 0.05}$	-	-	-	**.095**$_{\pm 0.06}$
	large	seen	-	-	.955$_{\pm 0.01}$	.969$_{\pm 0.01}$	.968$_{\pm 0.02}$	.920$_{\pm 0.02}$	.922$_{\pm 0.01}$	-.002$_{\pm 0.02}$	.912$_{\pm 0.01}$
		unseen	.440$_{\pm 0.03}$	-	-	.457$_{\pm 0.03}$	.419$_{\pm 0.01}$	.970$_{\pm 0.02}$	.893$_{\pm 0.02}$	**.077**$_{\pm 0.03}$	.876$_{\pm 0.04}$
		Δ	-	-	-	**.512**$_{\pm 0.03}$	**.549**$_{\pm 0.02}$	-	-	-	**.036**$_{\pm 0.04}$
RoBERTa	base	seen	-	-	.914$_{\pm 0.02}$	.898$_{\pm 0.03}$	.851$_{\pm 0.07}$	.832$_{\pm 0.03}$	.754$_{\pm 0.08}$	.078$_{\pm 0.09}$	.749$_{\pm 0.06}$
		unseen	.468$_{\pm 0.03}$	-	-	.388$_{\pm 0.02}$	.318$_{\pm 0.02}$	.846$_{\pm 0.04}$	.677$_{\pm 0.12}$	**.169**$_{\pm 0.13}$	.669$_{\pm 0.05}$
		Δ	-	-	-	**.510**$_{\pm 0.04}$	**.533**$_{\pm 0.07}$	-	-	-	**.080**$_{\pm 0.08}$
	large	seen	-	-	.937$_{\pm 0.03}$	.970$_{\pm 0.01}$	.984$_{\pm 0.01}$	.914$_{\pm 0.03}$	.907$_{\pm 0.01}$	.007$_{\pm 0.03}$	.819$_{\pm 0.13}$
		unseen	.460$_{\pm 0.02}$	-	-	.445$_{\pm 0.03}$	.399$_{\pm 0.04}$	.967$_{\pm 0.02}$	.884$_{\pm 0.01}$	**.083**$_{\pm 0.02}$	.799$_{\pm 0.11}$
		Δ	-	-	-	**.525**$_{\pm 0.03}$	**.585**$_{\pm 0.04}$	-	-	-	**.020**$_{\pm 0.17}$
XLM-RoBERTa	base	seen	-	-	.768$_{\pm 0.05}$	.683$_{\pm 0.01}$	.649$_{\pm 0.02}$	.690$_{\pm 0.09}$	.607$_{\pm 0.02}$	.083$_{\pm 0.09}$	.553$_{\pm 0.06}$
		unseen	-	.411$_{\pm 0.03}$	-	.238$_{\pm 0.01}$	.309$_{\pm 0.02}$	.678$_{\pm 0.06}$	.541$_{\pm 0.01}$	**.137**$_{\pm 0.06}$	.553$_{\pm 0.06}$
		Δ	-	-	-	**.445**$_{\pm 0.01}$	**.340**$_{\pm 0.03}$	-	-	-	**.000**$_{\pm 0.08}$
	large	seen	-	-	.941$_{\pm 0.01}$	.952$_{\pm 0.02}$	.955$_{\pm 0.03}$	.883$_{\pm 0.05}$	.862$_{\pm 0.06}$	.021$_{\pm 0.08}$	.761$_{\pm 0.08}$
		unseen	-	.488$_{\pm 0.03}$	-	.455$_{\pm 0.04}$	.383$_{\pm 0.02}$	.935$_{\pm 0.06}$	.783$_{\pm 0.08}$	**.152**$_{\pm 0.10}$	.735$_{\pm 0.09}$
		Δ	-	-	-	**.497**$_{\pm 0.04}$	**.572**$_{\pm 0.04}$	-	-	-	**.026**$_{\pm 0.12}$

Table 8: Results on our test data (average accuracy and standard deviation of five runs).

tion 3.4, as well as on all of the training data.

In all experiments, we conduct five trials and calculate the averages and standard deviations of the accuracy of the models. Training details are provided in Appendix F.

5 Results and Discussion

Table 8 shows the results of all our experiments. Overall, monolingual models with larger model sizes tend to perform better. In this section, we describe the results for each setting in detail.

5.1 Zero-shot setting

The two left columns in Table 8 show the results on the zero-shot setting. As Table 8 shows, the accuracy of both the monolingual and cross-lingual models is approximately 40%, and there is no significant difference between them. One possible reason is that SNLI, SICK, and their Japanese versions (JSNLI and JSICK) do not contain temporal inference, and the temporal inference patterns obtained from PLMUTE are only a fraction of the inference patterns required to solve our test set.

5.2 Fine-tuning setting

The right side of Table 8 shows the results on the fine-tuning setting. As expected, all models are highly accurate on the IID split setting (i.e., the setting in which all training data were used). We then discuss the results of the experiments using the splits described in Section 3.4.

Tense Fragment-based Split In the tense fragment-based split, the difference in accuracy between seen and unseen problems was nearly 50% for all models on both FRAGMENT_EASY and FRAGMENT_HARD. This suggests that the models cannot generalize the temporal inferences obtained from the training data.

Table 9 shows an example of unseen problems that RoBERTa-large could not solve on FRAGMENT_EASY and the corresponding seen problems in the training data. Because all models obtained similar results in relation to the generalization ability of LMs for temporal inference, we focus on the RoBERTa-large model, which achieved the best performance on our dataset. For this example, the model gave the same prediction for the both unseen and seen problems. The other tense fragment problems that the model could not solve on FRAGMENT_EASY have the same characteristics. Specifically, the model tended to predict incorrect labels for problems in which the premises and hypotheses of seen and unseen problems were very similar (differences are highlighted in bold), but the gold labels were different, as shown in Table 9. This suggests that this model does not capture the essential meaning of a sentence but determines the entailment relations based only on superficial information (i.e., the model does not generalize temporal inference patterns).

Time Format-based Split As shown in Table 8 shows, all models except XLM-RoBERTa-base achieved 80% accuracies on both unseen problems and seen problems of FORMAT_EASY. Furthermore, detailed analysis revealed that the XLM-RoBERTa-base did not solve problems that required inference of the size relationships between time units. This indicates that XLM-RoBERTa-base only fails to generalize the size relation between time units. One potential reason for this is that this model is cross-lingual and not large. In contrast, on FORMAT_HARD, all models exhibited reduced accuracy for the unseen problems compared to the seen problems. This indicates that the models do not have a priori knowledge regarding the size relationships between time units. There-

	Seen problem	Unseen problem
	TF: Habituality - Unmentioned TP + Always Gold label: Neutral	TF: Habituality + Negation - Unmentioned TP + Always Gold label: Contradiction, Pred label: Neutral
P	イヴァン は いつも 図面 を 遅れて 出す。 Ivan wa always drawing o late submit . (Ivan always submits his drawing late.) 2011 年 11 月 28 日 16 時 に イヴァン は 図面 を 出した。 2011 year 11 month 28 day 16 hour ni Ivan wa drawing o submit . (Ivan submitted his drawing at 16:00 on November 28, 2011.)	デイヴ は いつも マンション を 遅れて 訪れる。 Dave wa always apartment o late visit . (Dave always visits the apartment late.) 2002 年 5 月 11 日 14 時 に デイヴ は マンション を 訪れた。 2002 year 5 month 11 day 14 hour ni Dave wa apartment o visit . (Dave visited the apartment on May 11, 2002 at 14:00.)
H	イヴァン は 2011 年 11 月 28 日 22 時 に 図面 を 遅れて 出した。 Ivan wa 2011 year 11 month 28 day 22 hour ni drawing o late submit . (Ivan submitted his drawing late at 22:00 on November 28, 2011.)	デイヴ は 2012 年 2 月 1 日 0 時 に マンション を **遅れ ず** に 訪れた。 Dave wa 2012 year 2 month 1 day 0 hour ni apartment o late **not** ni visit . (Dave visited the apartment on February 1, 2012 at 0:00 **without** delay.)

Table 9: An example of unseen problem that RoBERTa-large could not solve in FRAGMENT_EASY and the corresponding seen problem in the training data. TF means the tense fragment.

fore, we consider that on FORMAT_EASY, BERT and RoBERTa succeeded in generalizing the inference patterns of the size relationships between time units based on minimal combinations of time units in the training data.

Time Span-based Split On the time span-based split, the large models achieved comparable accuracy on both the seen and unseen problems, whereas the base models tended to exhibit lower accuracy on the unseen problems. This suggests that the large models can generalize methods for determining the order relationships between time points, but the base models cannot generalize.

6 Conclusion

In this study, we constructed JAMP, a temporal Japanese NLI dataset, using a template-based approach. Our dataset is controllable in terms of difficulty, vocabulary, and size based on this approach. We conducted experiments using our dataset to probe the generalization ability of pre-trained language models for temporal inference. The experimental results indicated that current LMs can generalize for time format splits and time span splits but fail to generalize for tense fragment splits. Our dataset demonstrates that there is room for improvement in the generalization ability of current standard LMs for temporal inference. Because our method is applicable to the construction of datasets for other linguistic phenomena (e.g., modality, comparative), we plan to investigate the generalization ability of language models for other phenomena using the template-based approach in the future.

7 Limitations

In this section, we discuss two limitations of this study. The first limitation is that aspect and temporal commonsense are outside the scope of our dataset. Here, temporal commonsense refers to knowledge regarding events and the appropriate duration of those events. For example, the event "I washed my face for three years" is unnatural in terms of temporal commonsense, but this study did not consider such unnaturalness.

The second limitation is that the proposed method is currently applicable only to Japanese. In this study, we used a Japanese case frame dictionary to generate natural sentences. However, other languages such as English do not have resources equivalent to such a dictionary. Therefore, to apply our method to additional languages, we must first prepare a case frame dictionary for each language.

Acknowledgements

We thank the two anonymous reviewers for their helpful comments and suggestions, which improved this paper. This work was supported by JST, PRESTO grant number JPMJPR21C8, Japan.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Zeming Chen and Qiyue Gao. 2022. Curriculum: A broad-coverage benchmark for linguistic phenomena in natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3204–3219, Seattle, United States. Association for Computational Linguistics.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, Manfred Pinkal, David Milward, Massimo Poesio, Stephen Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework. Technical Report LRE 62-051r, The FraCaS Consortium.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA. Association for Computational Linguistics.

Ai Kawazoe, Ribeka Tanaka, Koji Mineshima, , and Daisuke Bekki. 2015. An inference Problem Set for Evaluating Semantic Theories and Semantic Processing systems for Japanese. In *JSAI International Symposium on Artificial Intelligence*, volume Technical report LRE 62-051r, FraCaS Consortium. Springer.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

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, Reykjavik, Iceland. European Language Resources Association (ELRA).

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, Florence, Italy. Association for Computational Linguistics.

Arindam Mitra, Ishan Shrivastava, and Chitta Baral. 2020. Enhancing natural language inference using new and expanded training data sets and new learning models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8504–8511.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8713–8721.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon. *Ph. D. Thesis, University of Pennsylvania*.

Tomoki Sugimoto and Hitomi Yanaka. 2022. Compositional semantics and inference system for temporal order based on Japanese CCG. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 104–114, Dublin, Ireland. Association for Computational Linguistics.

Shivin Thukral, Kunal Kukreja, and Christian Kavouras. 2021. Probing language models for understanding of temporal expressions. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 396–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

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

A Tense Fragment

Table 10 shows the tense fragments we defined.

Tense Fragment	Sub-tense Fragment
Temporal commonsense	Usage of 現在 (*now*)
Temporal ordering	Continuity of state
	Ordering relation
Time point	Mentioned time point
	Unmentioned time point
Temporal anaphora	Reference resolution of 昨日 (*yesterday*)
Interval	Comparison of two intervals
	Completion of eventuality
Habituality	Mentioned time point
	Unmentioned time point
	Negation
	Existential quantification

Table 10: Tense fragments we introduced in this study.

B Problem Creation for Some JSeM Problems

Table 11 shows examples of created problems and corresponding original problems in JSeM. As shown in Table 11, original and new problems are similar but have different gold labels. We also create templates for these created problems.

C Temporal Expression Generation in SHORT Time Span

The temporal expressions in SHORT are generated as follows. In the case of generating intervals, they are generated as described in Section 3.2, except that the integer selection range is one to three instead of one to nine. In the case of generating time points, we first identify the next largest unit after the smallest unit of the time format in the current problem and then calculate the duration of one-third of that unit. We then determine a selection range from a randomly selected time point to a time point that is advanced by the calculated duration. For example, if the smallest unit is "hour," then the next smallest unit is "day," so the selection range is between a specific time point and another time point one-third of a day (eight hours) in the future.

D Details for Dataset Artifacts Analysis

As mentioned in Section **??**, dataset artifacts analysis reveals correlations between labels and specific words. Formally, this analysis is a one-side binomial hypothesis test with the null hypothesis $p(y|x_i) = 1/3$, where $y \in \{Entailment, Neutral, Contradiction\}$, and x_i is a

	Original problem	New problem
	Gold label: Entailment	Gold label: Contradiction
P	スミス は ジョーンズ が 去る 前 に 去った。 Smith wa Jones ga leave before ni leave . (Smith left before Jones left.) ジョーンズ は アンダーソン が 去る 前 に 去った。 Jones wa Anderson ga leave before ni leave . (Jones left before Anderson left.)	スミス は ジョーンズ が 去る 前 に 去った。 Smith wa Jones ga leave before ni leave . (Smith left before Jones left.) ジョーンズ は アンダーソン が 去る 前 に 去った。 Jones wa Anderson ga leave before ni leave . (Jones left before Anderson left.)
H	スミス は アンダーソン が **去る 前** に 去った。 Smith wa Anderson ga leave **before** ni leave . (Smith left **before** Anderson left.)	スミス は アンダーソン が **去った 後** に 去った。 Smith wa Anderson ga leave **after** ni leave . (Smith left **after** Anderson left.)
	Gold label: Neutral	Gold label: Entailment
P	スミス が 2 時間 **以内 に** 報告書 を 書いた。 Smith ga 2 hour **within** ni report o write . (Smith wrote a report **within** two hours.)	スミス が 2 時間 **で** 報告書 を 書いた。 Smith ga 2 hour **de** report o write . (Smith wrote a report **in** two hours.)
H	スミス は その 報告書 を 書く の に 2 時間 を 費やした。 Smith wa that report o write no ni 2 hour o spent . (Smith spent two hours writing that report.)	スミス は その 報告書 を 書く の に 2 時間 を 費やした。 Smith wa that report o write no ni 2 hour o spent . (Smith spent two hours writing that report.)

Table 11: Examples of created problems and corresponding original problems in JSeM.

Section	Size
Train	9,750 (3,050/3,340/3,360)
Test	344 (114/112/118)

Table 12: JAMP dataset statistics. The lower row in parentheses shows the number of entailment, contradiction, and neutral examples, respectively.

Dataset Name	Size
SNLI (Bowman et al., 2015)	550,152
SICK (Marelli et al., 2014)	9,840
PLMUTE (Thukral et al., 2021)	72,720
JSNLI (Yoshikoshi et al., 2020)	533,005
JSICK (Yanaka and Mineshima, 2022)	5,000
PLMUTE_ja (Sugimoto and Yanaka, 2022)	11,220

Table 13: Statistics of dataset used in our experiments

word included in the vocabulary. For this analysis, we first split the hypothesis and premise sentences into individual words/tokens using Juman++ (Morita et al., 2015). We then count the number of occurrences of the gold label y in the n_i examples for every word x_i present in those examples. $p(y|x_i)$ is estimated based on the fraction of the count of the gold label y over n_i. According to the protocol described in Gardner et al. (2021), the null hypothesis is either accepted or rejected with a significance level of $\alpha = 0.01$ based on the Bonferroni correction.

E Data Statistics

Table 12 shows JAMP dataset statistics. Table 13 shows sizes of datasets used in our experiments.

F Training Details

We select the best learning rate among [6e-6,8e-6,1e-5,1.2e-5,2e-5] based on the development set. We use a batch size of 16 for training and eight for test.

G Data Licensing

Japanese case frame dictionary is distributed by Gengo-Shigen-Kyokai. JSeM is licensed under by BSD-3-Clause license. Our use of these two datasets is consistent with the terms of the license.