Enhancing the JNLI Dataset and Evaluating Model Performance on Improved Data

Jun Liang　Hiroaki Fujimoto　Masahiro Fukuyori
Fujitsu Limited, AI Laboratory, Japan
{liang-jun, fujimoto.hiroak, fukuyori}@fujitsu.com

Abstract

The Japanese Natural Language Inference (JNLI) dataset is a valuable resource for NLI research. However, we found it contains inconsistencies and lacks structural diversity. This paper presents a two-pronged approach to address these limitations: A rigorous correction of errors and the creation of a new, expanded dataset with diverse sentence structures. We detail our iterative correction methodology, leveraging Large Language Model (LLM) predictions and manual review. The new dataset introduces variations in sentence type (noun, verb, adjective/quantifier), enriching the data. Furthermore, we evaluate the performance of our internally created LLM model Takane on the original, corrected, and newly created JNLI datasets, demonstrating superior performance compared to existing state-of-the-art models.

1 Introduction

Natural Language Inference (NLI) remains a challenging yet crucial task in Natural Language Processing (NLP). In a NLI task, the goal is to determine the semantic relationship between a pair of sentences: a premise and a hypothesis. Premise: This is the given sentence, the statement that provides context or background information. Think of it as the established fact or assertion. Hypothesis: This is the sentence that needs to be evaluated in relation to the premise. It is a claim or statement that is being tested against the premise. The task is to determine the relationship between the premise and the hypothesis. This relationship is typically categorized into one of three classes/labels: Entailment: The hypothesis is logically implied by the premise. In other words, if the premise is true, the hypothesis must also be true. There is a clear logical connection. Contradiction: The hypothesis directly contradicts the premise. If the premise is true, the hypothesis must be false. They are opposing statements. Neutral: There is no clear logical relationship between the premise and the hypothesis. The truth of one doesn't necessarily affect the truth of the other. They are independent statements.

The creation and generation of high-quality, diverse datasets is helpful to evaluate a LLM's performance on logical thinking. The Japanese Natural Language Inference (JNLI) dataset[1] serves as an important resource for Japanese NLI research. However, its limitations, including inconsistencies and a lack of structural diversity, necessitate improvements. This paper addresses these limitations by presenting:

- A refined JNLI dataset through error correction;
- A novel, expanded JNLI dataset with diverse sentence structures.

We then evaluate the performance of Takane on these improved datasets and compare its performance against existing state-of-the-art models, demonstrating the effectiveness of our data enhancement strategies.

2 Methodology

In this section, data correction and data creation are written as below:

2.1 Dataset Correction

Of 88 randomly selected original JNLI ground truth (GT) sentence pairs reviewed by human annotators, only 24 (about 27%) were judged logically correct; the remainder were deemed incorrect or inappropriate. Thus, our dataset correction process begins by identifying inconsistencies in the original JNLI validation set. Originally, a sentence pair is chosen as GT if 6 out of 10 annotators give the same label. In our procedure, we ask 5 annotators to manually check the sentence pairs without telling them the GT in the

old dataset. We try to keep the fairness of the judgement and identify the types of the potential errors by setting different judge agreement rules. We represent three different confidence percentages for different inference labels: entailment, neutral, contradiction. 60% judge agreement: means for one sentence pair, 3 out of 5 annotators have the same judgement. 80% judge agreement: means for one sentence pair, 4 out of 5 annotators have the same judgement. 100% judge agreement: means for one sentence pair, 5 out of 5 annotators have the same judgement.

Firstly, we conform to the original paper[2] preparation procedure, finding that the sentence pair is made as the caption of a photo. The annotators are asked to give sentence pairs which can be used to describe this photo. Besides that, the relation between the two sentences in the sentence pair should be one of the three different inference labels: entailment, neutral, contradiction.

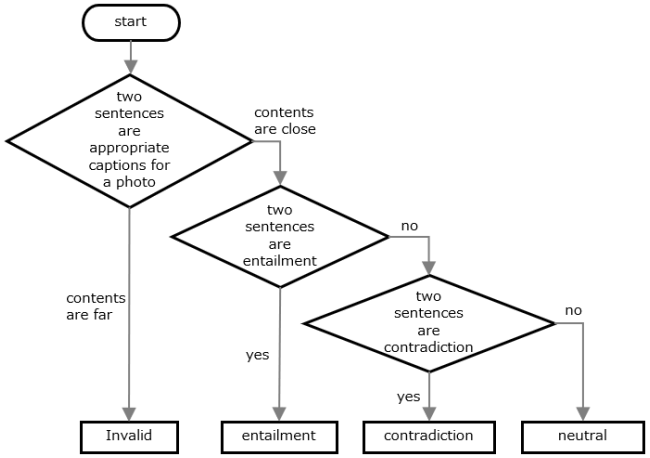

Figure 1: The review flow to determine the validity and label of a sentence pair

Figure 1 depicts a review process for evaluating the appropriateness of two sentences as captions for a photo. The process begins by determining if the contents of the two sentences are "close" enough to be considered appropriate captions for the same photo. Then, we sequentially determine the appropriate label. Please check the Appendix A.1 for more details.

2.2 Dataset Creation

The original work only asked annotators to describe the photo without any prerequisite. To address the lack of structural diversity in the original JNLI dataset, we create a new expanded JNLI dataset. This dataset focuses on expanding the variety of sentence structures. For each premise, we generate three hypothesis sentences, each exhibiting different types of transformation: noun, verb, and adjective/quantifier. Each hypothesis is carefully crafted to represent one of the three types of transformation. Please check the Appendix A.2 for more details. The creation of this dataset aims to provide a more challenging and representative benchmark for LLM training, avoiding overly simplistic examples that could lead to inflated performance metrics.

2.2.1 Basic Policy

The basic policies of the newly created dataset creation are:

- Conform to the original paper preparation procedure
  The sentence pair is given as the caption of a photo, and three different transformations of sentences are created without the photo.
- Wide variety
  Create text with three different transformations on one source text.
- Ensuring data quality
  To ensure data quality and prevent overfitting to easily processed data, we instruct annotators to create a more complex dataset that challenges even human judgment at first glance. Learning on complex data can also be very difficult, but complex is good for evaluation.

2.2.2 Creation Procedure

The new dataset creation procedure is as shown below:

1. Create a new premise that can be used as a caption for a photo which is not from the original text.
2. Create three hypotheses following the new premise, one for each of entailment, contradiction, or neutral according to the instructions and ideas on the following creation flow in Figure 1. The newly created sentence pair is totally different from the originally provided sentence pair on content.
3. Enhance each hypothesis with three transformations (noun, verb, adjective/quantifier) for entailment, contradiction, or neutral.
4. Check if the created data label: Entailment, Contradiction, and Neutral is correct by validating it again through the flow in Figure 1. Thus, there are nine

premise-hypothesis sentence pairs for one photo.

Please check the Appendix A.2 for the details.

3 Experimental Setup and Results

We evaluate our Takane model with the current state of art models on the original dataset, corrected dataset, and newly created dataset. We also evaluate the influence of differences in judge agreements on inference results.

3.1 Models Accuracy on Three Datasets

To evaluate the impact of our dataset enhancements, we conducte an experiment on these three datasets. Meanwhile, we also compare Takane model with current state-ot-art models.

The models are evaluated on the original JNLI validation set, correctd JNLI valication dataset, and newly created JNLI dataset.

Model	Original	Corrected	New
Takane	**0.890**	**0.923**	**0.770**
Command-R-plus-08-2024	0.689	0.693	0.672
Command-R-plus	0.644	0.645	0.626
GPT-4-0613	0.832	0.840	0.696
GPT-4	0.835	0.840	0.698
GPT-3.5	0.816	0.840	0.709
GPT-3.5-Turbo	0.719	0.719	0.663

Table 1: Model accuracy on these three types of datasets

GPT series models are based on underlying principles [3] and [4]. Command R+ models are from https://docs.cohere.com/v2/docs/command-r-plus. Takane is described in https://pr.fujitsu.com/jp/news/2024/09/30.html. The accuracy for each model is reported in Table 1. To ensure a fair comparison, the "Corrected Dataset" results use 80% annotator judge agreement— a more robust threshold rather than the original 60%. This decision is supported by our experiments on Judge Agreement in the next section. The Takane model outperforms all other models.

3.2 Effect of Different Judge Agreements

Since Takane reaches the best performance on the corrected datasets, we also evaluate the model on the newly created dataset with three different annotator judge agreements.

This experiment also shows that Takane performs the best among all the models.

Model	60%	80%	100%
Takane	**0.878**	**0.923**	**0.945**
Command-R-plus-08-2024	0.667	0.693	0.711
Command-R-plus	0.627	0.645	0.664
GPT-4-0613	0.807	0.840	0.877
GPT-4	0.807	0.840	0.878
GPT-3.5	0.807	0.840	0.877
GPT-3.5-Turbo	0.686	0.719	0.767

Table 2: Model accuracy on different judge agreements

Furthermore, we evalute the precision, recall, and F1-score of "entailment", "neutral", "contradiction" in each judge agreement.

As shown in Figure 2, precision, recall, F1-score, and accuracy all increase with higher annotator judge agreement (60%, 80%, 100%). However, the performance improvement from 60% to 80% annotator judge agreement is greater than the improvement from 80% to 100%. Therefore, 80% annotator judge agreement represents the optimal balance between data quality and performance gains.

3.3 Analysis Ratios of Inference Labels

We also analyze the ratios the three inference labels in each judge agreement in the corrected dataset, finding that the "neutral" label takes up roughly 50% of the valid dataset according to the result in Figure 3. To equally evaluate the labels of the inference, it is better to make a balanced dataset to validate the performance of the models.

4 Discussion

The results presented in Table 1 and Table 2 reveal a significant impact of data quality and structural diversity on the performance of NLI models. Takane consistently outperforms state-of-the-art models across all three datasets, demonstrating the effectiveness of our data enhancement approach. The augmentation in accuracy observed when moving from the original to the corrected dataset (Table 1) quantifies the negative impact of inconsistencies in the original JNLI dataset. This rise highlights the importance of addressing these errors for reliable model evaluation. Furthermore, the performance difference between the corrected and the new dataset demonstrates the significant impact of structural diversity. The introduction of varied sentence structures, as detailed in Section 2.2, creates a

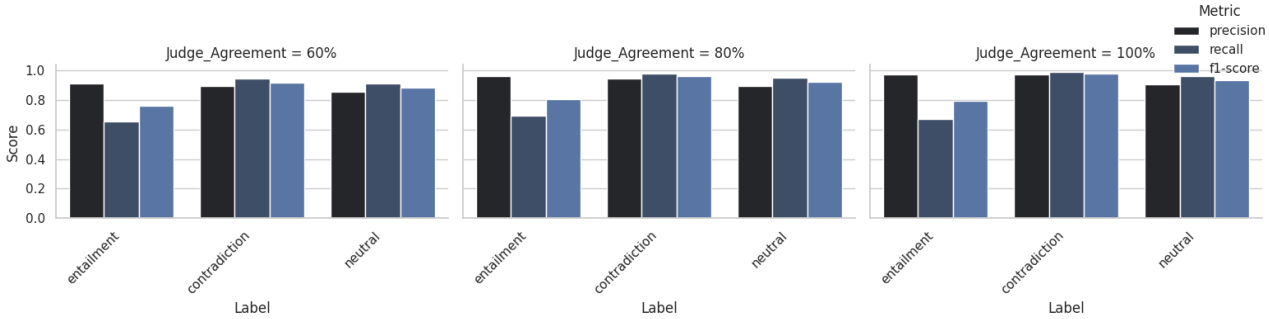

Figure 2: Takane model's Precision, Recall, F1-score among different judge agreements

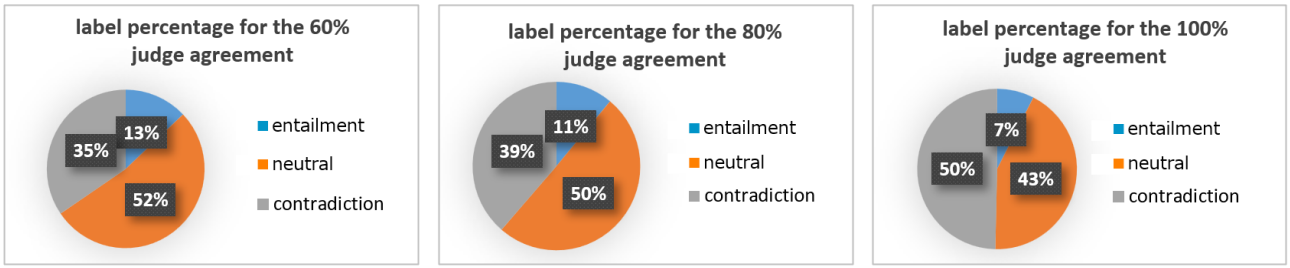

Figure 3: Inference label ratios analysis across different annotator judge agreements

more challenging and realistic benchmark for evaluating NLI models. The fact that all models' performance decreases on the structurally diverse dataset, while Takane is still outperforming other models, suggests that current state-of-the-art models still struggle with more complex syntactic structures.

The analysis of different inter-annotator judge agreement levels (Table 2) reinforces this observation. The consistent increase in accuracy judging by Table 2 with increasing agreement thresholds (from 60% to 100%) is expected, as higher agreement levels indicate more nuanced and challenging examples. However, Takane consistently outperforms other models at all these three agreement levels, demonstrating its robustness. The detailed precision, recall, and F1-score analysis (Figure 2) further illuminates the model's strengths and weaknesses across different labels at each agreement level.

The analysis of the original valid dataset inference label distribution reveals a notable class imbalance, with neutral examples consistently comprising approximately 50% of the dataset across all agreement levels. This imbalance may influence model evaluation, potentially leading to inflated performance metrics for the other certain inference labels.

5 Conclusion

This study demonstrated the crucial role of data quality and structural diversity in advancing Japanese NLI. Our two-pronged approach for rigorous error correction and the creation of a structurally diverse expanded dataset significantly improved the JNLI dataset. Both data correction and data creation resulted in superior performance for our Takane model compared to existing state-of-the-art models. Analysis revealed the impact of data inconsistencies and label imbalance.

Future research should focus on several key areas. First, expanding the dataset further, particularly addressing the label imbalance and incorporating a wider range of sentence structures, is crucial. Second, investigating alternative error correction methods and exploring the use of active learning techniques for dataset expansion could further enhance data quality. Third, evaluating the transferability of our improved dataset to other downstream NLP tasks is essential to assess its broader impact. Finally, a deeper linguistic analysis of the factors contributing to the difficulty of Japanese NLI tasks will inform the design of future datasets and models, leading to more robust and reliable systems for Japanese language understanding.

Acknowledgements

We are extremely grateful to Shigeyuki Odashima and Susumu Tokumoto for their consistent support throughout this research. We also thank Shigeyuki Odashima and Pelat Guillaume for their proofreading.

References

[1] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, Marseille, France, 2022. European Language Resources Association.

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015.

[3] Tom B. Brown et al. Language models are few-shot learners, 2020.

[4] OpenAI et al. Gpt-4 technical report, 2024.

A Appendix

The details of data correction and data creation are as shown below.

A.1 Data Correction Process

In this section, review flowchart is described to better help understand data correction process.

- If the contents are close: The process moves to check for entailment. Entailment means one sentence logically implies the other. We have created entailment check steps (the details will be shown in the dataset review flow: Figure 1). If the sentence pair is judged as "entailment", then the flow stops. If the sentence pair is judged as "not entailment", the process then checks for "contradiction". Contradiction means the sentences oppose each other. The process follows the contradiction check steps. If the contradiction is found, the result is "contradiction"; if not, the final output is "neutral".
- If the contents are far: The process determines the sentences are inappropriate as datasets, resulting in the output "Invalid" as data inappropriation. In short, the flowchart outlines a decision tree for classifying sentence pairs (intended as image captions) into one of four labels: entailment, contradiction, neutral, or invalid. The core logic relies on assessing the semantic closeness of the sentences and then applying specific rules for entailment and contradiction determination. The determination steps on entailement and contradiction will be shown in Figure 1.

A.2 Data Creation Process

This section details the creation of the expanded JNLI dataset, focusing on increased structural diversity to create more challenging examples for NLI model evaluation. The process adhered to these principles:

1. **Adherence to Original Methodology:** Sentence pairs are created as if they are photo captions, but without providing the actual photos.
2. **Structural Diversity:** Three sentence transformations (noun, verb, adjective/quantifier) are applied to each premise.
3. **Data Quality Assurance:** Annotators are instructed to create complex examples challenging even human judgment.

The procedure of data creation is described in 2.2.2.

A.2.1 Data Creation Example

For one original sentence pair, we expand it into three labels. Three transformations are made for each label. The newly created premise should be totally different from the original premise.

Entailment

- Noun: Premise: 猫が芋虫にじゃれています。Hypothesis: 生き物が芋虫にじゃれています。
- Verb: Premise: 猫が物陰からネズミを狙っています。Hypothesis: 猫がネズミを見ています。
- Adjective/Quantifier: Premise: 映画館で男女のペアが話をしている。Hypothesis: 映画館で夫婦が話をしている。

Contradiction

- Noun: Premise: 公園で女の子がボール遊びをしている。Hypothesis: 公園で少年がボール遊びをしている。
- Verb: Premise: 公園で女の子がかけっこをしている。Hypothesis: 公園で女の子が泣いている。
- Adjective/Quantifier: Premise: 公園で幼い女の子がブランコに乗っている。Hypothesis: 公園で老婆がブランコにのっている。

Neutral

- Noun: Premise: 映画館で男女のペアが話をしている。Hypothesis: 映画館で夫婦が話をしている。
- Verb: Premise: 映画館で夫婦が話をしている。Hypothesis: 映画館で夫婦がけんかをしている。
- Adjective/Quantifier: Premise: 映画館で若いカップルがポップコーンを買っている。Hypothesis: 映画館で若いカップルがキャラメル味のポップコーンを買っている。