IndoJavaneseNLI: A Cross-Lingual Natural Language Inference Dataset for East Javanese "Ngoko" Register

Jalaluddin Al-Mursyidy Fadhlurrahman[1], Ayu Purwarianti[2]
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
23521059@std.stei.itb.ac.id[1], ayu@staff.stei.itb.ac.id[2]

Alham Fikri Aji
Natural Language Processing Department
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirate
alham.fikri@mbzuai.ac.ae

***Abstract*—Natural Language Inference (NLI) is a task that focuses on establishing the logical relationship between two sentences, premise sentences, and hypothesis sentences, by classifying it into "entailment", "neutral", and "contradiction". The growth of NLI models has been noteworthy in the English language. However, little to no progress has been made on low-resource languages, such as Javanese. What's more, the resulting translation from Machine Translation is blind to the cultural nuance of word choices. It tends to mix up words from various Javanese registers like "Ngoko", "Madya", and "Krama". To tackle that issue, we present IndoJavaneseNLI, a cross-lingual NLI dataset for East Javanese "Ngoko" registers. Our dataset consists of the premise sentences in the Indonesian language and the hypothesis sentences in the Javanese language. This paper describes how we carefully inspect the Javanese translation, the process leading up to the building of good East Javanese "Ngoko" sentences, and its evaluation with Transformer-based models and transfer learning. We also found that XLMR yields the best result in solving NLI problems with our dataset with the accuracy of 67.56% on the fine-tuning method and 47.34% on the transfer learning method.**

***Index Terms*—natural language inference, nli, cross-lingual nli, javanese nli, zero-shot, low-resource language, natural language processing, nlp**

I. Introduction

Natural Language Inference (NLI), is a fundamental task in Natural Language Processing (NLP) that focuses on determining the logical relationship between two sentences, such as: "entailment", "neutral", and "contradiction" [6]. The development of NLI models has gained significant attention in recent years, driven by its application in text summarization [9, 12], context understanding [17], and question answering [20]. While substantial progress has been made in NLI research, most efforts have been concentrated on high-resource languages, such as English. [10], leaving many low-resource languages, like Javanese, understudied.

This paper introduces the IndoJavaneseNLI, a cross-lingual natural language inference dataset for Javanese, to address the scarcity of NLI data in underrepresented languages. The Javanese language, spoken by more than 98 million people in Indonesia and the 21^{st} most spoken languages in the world [8], serves as a prominent example of such a language. Despite its significant speaker base, research in NLI for Javanese has been largely neglected, limiting the potential for developing natural language understanding systems.

IndoJavaneseNLI is constructed to facilitate research on NLI in Javanese and bridge the gap between NLI advancement in high-resource and low-resource languages. It includes diverse sentence pairs drawn from various domains and genres. Leveraging IndoJavaneseNLI, researchers can advance the state of NLI for low-resource languages and contribute to developing more inclusive and globally applicable natural language processing systems.

In this paper, we present the details of the IndoJavaneseNLI dataset, its construction process, the data collection methodology, and an evaluation of baseline models' performance on the dataset. We aim to encourage and facilitate further research in NLI for low-resource languages, promoting linguistic diversity and inclusion in natural language processing. The release of IndoJavaneseNLI represents a step toward ensuring that NLI becomes accessible and effective for a broader range of languages, cultures, and communities, ultimately advancing the goal of a more inclusive and globally representative NLP research landscape.

II. Related Work

The Standford NLI (SNLI) [3] was one of the pioneering contributions in the field of NLI. It consists of 570,152 diverse sentence pairs with human-annotated labels. SNLI provided a robust benchmark for NLI systems and played a pivotal role in the early development of neural models for this task. However, the sentences in SNLI derived from image captions, limiting the hypothesis sentences to be short and simple and restraining many critical phenomena like modality and temporal reasoning.

Another effort has been made for NLI research, such as the Multi-NLI (MNLI) [22]. Unlike SNLI, MultiNLI features a more diverse set of genres and writing styles, making it a valuable resource for NLI research. MNLI comprises matched and mismatched sentence pairs, providing a challenging test

bed for evaluating the robustness of NLI systems. Unlike its name, this dataset does not have the data parallel in another language. Another work has been made to extend the MNLI dataset with data in other languages, like XNLI.

XNLI [5] is a cross-lingual extension of MNLI to promote research in cross-lingual NLI. It covers 15 languages, including two low-resource languages, such as Swahili and Urdu. It serves as a crucial dataset for assessing the cross-lingual transferability of NLI models, encouraging the development of systems capable of handling multiple languages. One limitation of XNLI is that it lacks the cultural nuance of the target languages because of the translation process.

The efforts to build datasets for other languages have been made using machine translation (MT) on existing English NLI datasets [15]. Sometimes, it is coupled with fewer human-annotated data[2, 16]. Except for the Original Chinese Natural Language Inference (OCNLI) dataset built using the human-annotation method [11].

Currently, there are three NLI datasets in the Indonesian language, namely WReTE [19], INARTE [1], and IndoNLI [14]. Both WReTE and INARTE datasets only have a small number of sentence pairs, 400 pairs for WReTE and $\pm$1.5k pairs for INARTE, and use only two labels: "entailment" and "not-entailment" [14]. This makes IndoNLI the largest Indonesian NLI dataset, with $\pm$18k sentence pairs. IndoNLI was created using three genres: Wikipedia, news, and Web articles.

III. Dataset Construction

A. Data Source

To tackle the lack of Javanese language in the NLI task, we propose a cross-lingual NLI dataset that consists of the Indonesian-Javanese language. Our IndoJavaneseNLI is readily available in a public repository[1]. We are using IndoNLI as our base data since IndoNLI is the largest NLI dataset in the Indonesian language. We only keep the premise sentences in Indonesian and translate the hypothesis into Javanese. The Javanese variant we use for our dataset is the East Javanese "Ngoko" variant of the Javanese language. We compare the resulting translation from Google Translate, Chat GPT OpenAI API, and Mongosilakan as our MT systems. The annotators then assessed the translation results of the three MT systems.

B. Annotation Protocol

1) Annotator's Requirements: All annotators must be fluent in Indonesian and Javanese. We only choose annotators who are native residents of East Java since they are expected to be able to understand better the cultural implementation of the East Javanese "Ngoko" variant of the Javanese language.

2) Annotation Evaluation: Every data set consists of a premise sentence written in Indonesian and a pair of translated hypothesis sentences in Javanese. Each annotator was given the same parallel set of data. Then, the annotators marked the quality of the resulting hypothesis sentence from each MT with a 1-5 scale ("5" is the best, and "1" is the lowest). The order of appearance of data from each MT system is randomized for each row, and the order of appearance is kept secret from the annotator to ensure the objectivity of the annotation process. The annotators would check whether the translated sentences are free from grammatical errors, spelling, and punctuation. Moreover, the annotators would also check if the resulting translation changed the semantic relation between the premise and the hypothesis sentences.

Observing the information between the two sentences could determine the semantic relation between sentences. A pair of premise-hypothesis sentences can be said to be "entailment" if it can be concluded that the hypothesis sentence is true based on the information in the premise sentence. If it can be concluded that the hypothesis sentence is false based on the information in the premise sentence, then the sentence pair is "contradiction". Otherwise, it is "neutral" if the truth of the hypothesis sentence cannot be determined based on the information in the premise sentence or if there is not enough information in the hypothesis sentence.

3) Gold Label: The gold label is the main label that will be used to mark the quality of the translation of hypothesis sentences in Javanese. Firstly, to obtain the gold labels, we need to get the average mark of the annotated labels from each annotator grouped by its MT sources. Then, we calculate the average of each MT system from all annotators. Next, we decide which MT system has the highest mark as a translation baseline. Finally, we re-calculate the label average from all annotators for that selected MT. The resulting average for the best MT data will be chosen as the gold label. If a gold label for a pair of data is less than 3, the data is flagged as "broken" and should be fixed.

4) Data Fixing: The data flagged as "broken" should be fixed using several strategies by another independent expert annotator. The newly created hypothesis sentence in Javanese could be made by removing or inserting one or more words from the premise sentence. The independent annotator could also replace one or more words from the premise sentence with a synonym, antonym, hypernym, or hyponym; or he could paraphrase. Alternatively, the independent annotator could change the sentence structure, for example, from passive to active. The fixed sentence must be in Javanese and would be re-annotated following the annotation protocol mentioned earlier.

5) Resulting Corpus and Annotation: After the first round of annotation, we have Google Translate as the best average annotation score (4.15 out of 5), followed closely by Mongosilakan (4.08 out of 5) and ChatGPT API (3.31 out of 5). As a result, we use Google Translate as our main MT source. However, the resulting translations still have various Javanese registers, such as "Ngoko" and "Krama", mixed up in one sentence, as can be seen in Table I in words in bold. For instance, the word "nate" and "kanggo" are words in Central Javanese "Ngoko" register, meanwhile, the words "tansah" and "minangka" are in the "Kromo Inggil" Javanese register. This may happened because the current MT system does not have any knowledge regarding the cultural usage

[1]https://github.com/jalalAzhmatkhan/indojavanese-nli

Hypothesis in Indonesian	Hypothesis in Javanese (Before Annotation)	Hypothesis in Javanese (After Annotation)
Yuko Hara tidak pernah mengambil cuti.	Yuko Hara ora **nate** liburan. *(Yuko Hara never takes leave.)*	Yuko Hara ora tau cuti.
Teks berjalan selalu diposting untuk kisah nyata.	Running text **tansah** dikirim **kanggo** crita nyata. *(Running text is always posted for the real story.)*	Tulisan mlaku mesti digawe nang kisah nyata
Hillary Clinton maju sebagai calon Presiden.	Hillary Clinton mlaku **minangka** calon presiden. *(Hillary Clinton is running as a presidential candidate.)*	Hillary Clinton maju nyalon presiden

Table I: Sample of translated data straight from MT system (the "Before Annotation" column) and data that has been annotated into East Javanese "Ngoko" registers.

Classes	Train	Validation	Test
Entailment	3476	807	808
Neutral	3415	749	629
Contradiction	3439	641	764

Table II: IndoJavaneseNLI corpus statistics.

	Jaccard	LCS	New token rate
Entailment	6.13	3.93	66.4
Neutral	4.36	3.21	72.04
Contradiction	5.95	3.73	68.28

Table III: Word overlap between premise and hypothesis sentence in test split.

of the Javanese registers. The annotation process in this step was quite challenging because of the many registers in the Javanese language. Ridiculously, some MT systems improvise by inserting some English vocabulary into their translation, as shown in Table I. We ran another round of annotation after fixing the data with an annotation score lower than a threshold. The gold label selection process yields an average score of 4.05, with only six broken sentences on the first round of annotation. After fixing the broken sentences, all resulting translated data yields an average score of 4.1 out of 5 without any broken sentences.

Since our dataset derives from IndoNLI [14], our dataset has the exact class count per split, as shown in Table II. We see that the three classes are relatively balanced. The word overlap analysis shown in Table III suggests that our dataset has minimal lexical overlap between the Indonesian premise sentence and the Javanese hypothesis sentence, both ordered and unordered. While the Jaccard index and LCS similarity scored very low, the opposite can be said for the new token rate. The new token rate measures the percentage of hypothesis tokens that do not exist in premise tokens. The high new token rate suggests that our dataset has a highly diverse token generated from having the premise and hypothesis in different languages.

IV. Experiments

We experiment with several Transformer-based models: multilingual BERT, BERT [7], IndoBERT [21], and XLMR [4]. We used base and large architecture for BERT and XLMR, as for the others, we used the base architecture only. We tried several experiments like fine-tuning the model, zero-shot experiments by only training in the Indonesian language, and transfer learning. We tried various training scenarios to test our dataset.

A. Experiment scenarios and Hyperparameters

For the mBERT fine-tuning experiment, we use a learning rate of 3×10^{-6}, 6 epochs, a token max length of 256, and a batch size of 8. The BERT$_{(base)}$ and IndoBERT used the same hyperparameters, except for the batch size which is 4. Fine-tuning BERT$_{(large)}$ used learning rate 1×10^{-5}, 10 epochs, batch size 2, and a token max length of 512. We use learning rate 1×10^{-6}, token max length of 512, 15 epochs, and batch size 4 for fine-tuning the XLMR$_{(base)}$ model. To fine-tune the XLMR$_{(large)}$ model, we use learning rate 3×10^{-6}, token max length of 512, 6 epochs, and batch size 2. For our transfer learning method, we have the Transformer-based models fine-tuned in Indonesian language sentence pairs as the teacher models and use them to infer our IndoJavaneseNLI dataset.

We used the Huggingface Transformers [23] for all fine-tuning scenarios, AdamW [13] as an optimizer, and Pytorch [18] for all training experiments. Our research was done on various GPU devices, such as NVIDIA GeForce RTX 3060, NVIDIA A100 Tensor Core, and Tesla P100 Data Center Accelerator.

V. Results and Analysis

A. Fine-Tuning Transformer-based Models

Table IV reports the performance of fine-tuning the Transformer-based models. As expected, the zero-shot method yields the lowest accuracy even on the XLMR model, which is 44.51%. The other models yield an average of $60s\%$ in accuracy, except when using XLMR$_{(large)}$. The XLMR$_{(large)}$ model outperforms others, even IndoBERT. This may indicate that XLMR$_{(large)}$ has a larger training corpus in Javanese than the other models and architecture. This finding is in line with IndoNLI and IndoNLU benchmark results [14, 21].

B. Transfer Learning

Another experiment to test our dataset is by utilizing the transfer learning method. We fine-tuned the baseline models in the Indonesian language for both premise and hypothesis sentences and used the model to infer our dataset. Table V shows us that XLMR$_{(base)}$ yielded the highest score, outperforming the mBERT model.

Experiment	Accuracy (%)	F1 Score (%)
mBERT	62.9	63.1
IndoBERT$_{(base)}$	61.47	61.46
IndoBERT$_{(large)}$	62.92	63.06
BERT$_{(base)}$	60.63	60.88
BERT$_{(large)}$	59.83	60.13
XLMR$_{(base)}$	62.15	62.4
XLMR$_{(large)}$	67.56	67.7
XLMR$_{(zero-shot)}$	44.51	41.02

Table IV: Fine-Tuning experiment results using transformer-based models.

Experiment	Teacher	Student	Accuracy (%)	F1 Score (%)
mBERT	Indonesian - Indonesian	Indonesian - Javanese	43.28	40.15
XLMR$_{(base)}$	Indonesian - Indonesian	Indonesian - Javanese	47.34	45.55

Table V: Transfer learning experiment results.

VI. Conclusion

We present IndoJavaneseNLI, the NLI dataset specifically created for East Javanese "Ngoko" Registers. This dataset is created to address the lack of NLI data in Javanese. We found that the MT system does not fully understand the cultural nuance of word choices. The resulting translated sentences tend to mix up various Javanese registers, such as "Ngoko" and "Krama". To address this issue, we annotate the translated Javanese data and fix the word choices based on the annotation score. We have performed fine-tuning and transfer learning methods to evaluate existing language models for our dataset. We found that XLMR$_{(large)}$ yielded the best result in both experiments.

VII. Limitations and Ethical Considerations

We are using the IndoNLI [14] as our base, which was created using data from Wikipedia, various news, and web domains. The data origin of our source may contain harmful content and stereotypes.

Acknowledgements

We want to thank Hayyu Rachma Widya Aulya, Andika Octavia Pratama Putra, Muhammad Fahmy Nadhif, and Muhammad Daimus Suudi for their contribution in assisting the annotation process. We also thank Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI) for allowing us to use the NVIDIA A100 Tensor Core for parallel training.

References

[1] Abdiansah Abdiansah, Azhari Azhari, and Anny Kartika Sari. Inarte: An indonesian dataset for recognition textual entailment. In *2018 4th International Conference on Science and Technology (ICST)*, pages 1–5, 2018.

[2] Željko Agić and Natalie Schluter. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, September 2015. Association for Computational Linguistics.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[5] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[6] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Eberhard, M. David, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world, 2023. Accessed October 15th, 2023.

[9] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource

scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online, June 2021. Association for Computational Linguistics.

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

[12] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[14] Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[15] Yashar Mehdad, Matteo Negri, and Marcello Federico. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1345, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[16] Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 670–679, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[17] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[19] Ken Nabila Setya and Rahmad Mahendra. Semi-supervised textual entailment on indonesian wikipedia data. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 416–427, Cham, 2023. Springer Nature Switzerland.

[20] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics.

[21] Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China, December 2020. Association for Computational Linguistics.

[22] Adina Williams, Nikita Nangia, and Samuel Bowman. 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, June 2018. Association for Computational Linguistics.

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.