# IndoNLI: A Natural Language Inference Dataset for Indonesian

**Rahmad Mahendra**[1] **Alham Fikri Aji**[2] **Samuel Louvan**[3]
**Fahrurrozi Rahman**[4] **Clara Vania**[5*]
[1]Universitas Indonesia [2]Kata.ai Research [3]Fondazione Bruno Kessler
[4]University of St Andrews [5]Amazon
rahmad.mahendra@cs.ui.ac.id, aji@kata.ai, slouvan@fbk.eu,
fr27@st-andrews.ac.uk, vaniclar@amazon.co.uk



## Abstract

We present IndoNLI, the first human-elicited NLI dataset for Indonesian. We adapt the data collection protocol for MNLI and collect ~18K sentence pairs annotated by crowd workers and experts. The expert-annotated data is used exclusively as a test set. It is designed to provide a challenging test-bed for Indonesian NLI by explicitly incorporating various linguistic phenomena such as numerical reasoning, structural changes, idioms, or temporal and spatial reasoning. Experiment results show that XLM-R outperforms other pretrained models in our data. The best performance on the expert-annotated data is still far below human performance (13.4% accuracy gap), suggesting that this test set is especially challenging. Furthermore, our analysis shows that our expert-annotated data is more diverse and contains fewer annotation artifacts than the crowd-annotated data. We hope this dataset can help accelerate progress in Indonesian NLP research.

## 1 Introduction

Indonesian language or *Bahasa Indonesia* is the 10th most spoken language in the world with more than 190 million speakers.[1] Yet, research in Indonesian NLP is still considered under-resourced due to the limited availability of annotated public datasets. To help accelerate research progress, IndoNLU (Wilie et al., 2020) and IndoLEM (Koto et al., 2020) collect a number of annotated data to benchmark Indonesian NLP tasks.

In line with their effort, we introduce Indonesian NLI (INDONLI), a natural language inference dataset for Indonesian. Natural language inference (NLI), also known as *recognizing textual entailment* (RTE; Dagan et al., 2005) is the task of determining whether a sentence semantically entails another sentence. NLI has been used extensively as a benchmark for NLU, especially with the availability of large-scale English datasets such as the Stanford NLI (SNLI; Bowman et al., 2015) and the Multi-Genre NLI (MNLI; Williams et al., 2018) datasets. Recently, there have been efforts to build NLI datasets in other languages. The most common approach is via translation (Conneau et al., 2018; Budur et al., 2020). One exception is the work by Hu et al. (2020) which uses a human-elicitation approach similar to MNLI to build an NLI dataset for Chinese (OCNLI).

Until now, there are two Indonesian NLI datasets available. The first one is WReTE (Setya and Mahendra, 2018), which is created using revision history from Indonesian Wikipedia. The second one, INARTE (Abdiansah et al., 2018), is created automatically based on question-answer pairs taken from Web data. Both datasets have relatively small number of examples (400 pairs for WReTE and ~1.5k pairs for INARTE) and only uses two labels (*entailment* and *non-entailment*). Furthermore, since the hypothesis sentence is generated automatically from the premise, they tend to be so similar that arguably they will not be effective as a benchmark for NLU (Hidayat et al., 2021). On the other hand, INDONLI is created using a human-elicited approach similar to MNLI and OCNLI. It consists of ~18K annotated sentence pairs, making it the largest Indonesian NLI dataset to date.

INDONLI is annotated by both crowd workers (layperson) and experts. Lay-annotated data is used for both training and testing, while expert-annotated data is used exclusively for testing. Our goal is to introduce a challenging test-bed for Indonesian NLI. Therefore the expert-annotated test data is explicitly designed to target phenomena such as lexical semantics, coreference resolution, idioms expression, and common sense reasoning. Table 1 exemplifies INDONLI data.

*Work done while at New York University.

[1]https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world (Accessed May 2021).

| Premise | Hypothesis | Label | Phenomena |
|---|---|---|---|
| **LAY-ANNOTATED DATA** | | | |
| Seakan tak bisa dipisahkan, dua sahabat itu sama-sama sedang menggarap proyek musik.<br>*(As if they cannot be separated, the two friends are both working on music projects.)* | Dua sahabat itu selalu bersama-sama.<br>*(The two friends are always together.)* | N | n/a |
| Meskipun trikomoniasis adalah penyakit yang sangat umum, penyakit ini seringkali sulit diketahui.<br>*(Although trichomoniasis is very common disease, it is often difficult to detect.)* | Trikomoniasis bukanlah penyakit yang umum.<br>*(Trichomoniasis is not a common disease.)* | C | n/a |
| **EXPERT-ANNOTATED DATA** | | | |
| Selanjutnya, dua pemain Arsenal yang dirasa Ian Wright kurang sip adalah di sektor serang. Mereka adalah Willian dan Alexandre Lacazette , yang disebutnya buntu.<br>*(Furthermore, two Arsenal players who Ian Wright felt were not good enough were in the attack sector. They are Willian and Alexandre Lacazette , who he calls dead ends.)* | Alexandre Lacazette tidak memiliki performa yang baik sebagai penyerang.<br>*( Alexandre Lacazette did not have a good performance as an attacker. )* | E | MORPH , NEG , COREF , LEXSEM |
| Setelah dewasa, Ramlah dinikahi oleh Amr bin Utsman bin Affan.<br>*(After growing up, Ramlah was married to Amr bin Uthman bin Affan.)* | Ramlah lebih muda dari pada Amr bin Utsman.<br>*(Ramlah was younger than Amr bin Uthman.)* | N | COMP |

Table 1: Examples of premise and hypothesis pairs from INDONLI (**E:** Entailment, **C:** Contradiction, **N:** Neutral). English translation is provided in the bracket for context. The expert-annotated data is annotated with linguistic phenomena contributing to make the inference. For illustrative purposes, we highlight the sentence chunks that correspond to the specific phenomena (noting that such highlighting is not available in the released dataset).

We also propose a more efficient label validation protocol. Instead of selecting a consensus gold label from 5 votes as in MNLI data protocol, we incrementally annotate the label starting from 3 annotators. We only add more label annotation if consensus is not yet reached. Our proposed protocol is 34.8% more efficient than the standard 5 votes annotation.

We benchmark a set of NLI models, including multilingual pretrained models such as XLM-R (Conneau et al., 2020) and pretrained models trained on Indonesian text only (Wilie et al., 2020). We find that the expert-annotated test is more difficult than lay-annotated test data, denoted by lower model performance. The Hypothesis-only model also yields worse results on our expert-annotated test, suggesting fewer annotation artifacts. Furthermore, our expert-annotated test has less hypothesis-premise word overlap, signifying more diverse and creative text. Overall, we argue that our expert-annotated test can be used as a challenging test-bed for Indonesian NLI.

We publish INDONLI data and model at https://github.com/ir-nlp-csui/indonli.

## 2 Related Work

**NLI Data** Besides SNLI and MNLI, another large-scale English NLI data which is proposed recently is the Adversarial NLI (ANLI; Nie et al., 2020). It is created using a human-and-model-in-the-loop adversarial approach and is commonly used as an extension of SNLI and MNLI.

For NLI datasets in other languages, the Cross-lingual NLI (XNLI) corpus extends MNLI by manually translating sampled MNLI test set into 15 other languages (Conneau et al., 2018). The Original Chinese Natural Language Inference (OCNLI) is a large-scale NLI dataset for Chinese created using data collection similar to MNLI (Hu et al., 2020). Other works contribute to creating NLI datasets for Persian (Amirkhani et al., 2020) and Hinglish (Khanuja et al., 2020).

Some corpora are created with a mix of machine translation and human participation. The Turkish NLI (NLI-TR) corpus is created by machine-translating SNLI and MNLI sentence pairs into Turkish, which are then validated by Turkish native speakers (Budur et al., 2020). For Dutch, Wijnholds and Moortgat (2021) introduce the SICK-NL by machine-translating the SICK dataset (Marelli

et al., 2014). It is then manually reviewed to maintain the correctness of the translation.

AmericasNLI, an extension of XNLI to 10 indigenous languages of the Americas, is created with the primary goal to investigate the performance of NLI models in truly low-resource language settings (Ebrahimi et al., 2021). The dataset is built by translating Spanish XNLI into the target languages, and vice versa, using Transformer-based sequence-to-sequence models.

**NLI Analysis** Research conducted by Tsuchiya (2018); Poliak et al. (2018b); Gururangan et al. (2018) show that there are hidden biases in the NLI corpus, such as word choices, grammaticality, and sentence length, which allow models to predict the correct label only from the hypothesis.

Several studies also investigate whether NLI models might use heuristics in their learning. Many NLI models still suffer from various aspects such as antonymy, numerical reasoning, word overlap, negation, length mismatch, and spelling error (Naik et al., 2018), lexical overlap, subsequence and constituent (McCoy et al., 2019), lexical inferences (Glockner et al., 2018) and syntactic structure (Poliak et al., 2018a).

Research to analyze which linguistic phenomena are learned by current models has gained interest. This ranges from the definition of the diagnostic test (Wang et al., 2018), the linguistic phenomena (Bentivogli et al., 2010), fine-grained annotation scheme (Williams et al., 2020), to the taxonomic categorization refinement (Joshi et al., 2020).

## 3 IndoNLI Data Construction

### 3.1 Data Source

Our premise text is originated from three genres: Wikipedia, news, and Web articles. For the news genre, we use premise text from Indonesian PUD and GSD treebanks provided by the Universal Dependencies 2.5 (Zeman et al., 2019) and IndoSum (Kurniawan and Louvan, 2018), an Indonesian summarization dataset. For the Web data, we use premise text extracted from blogs and institutional websites (e.g., government, university, and school). To maximize vocabulary and topic variability, we set a limit of five text snippets from the same document as premise text. Moreover, the source of premise text covers a broad range of topics including, but not limited to, science, politics, entertainment, and sport.

| | Lay | | Expert | |
|---|---|---|---|---|
| Round | #pairs | cum-cons | #pairs | cum-cons |
| 1st | 4,489 | 74.3% | 3,008 | 80.0% |
| 2nd | 1,155 | 94.7% | 602 | 93.8% |
| 3rd | 237 | 98.0% | 188 | 99.2% |
| | #annotation needed (incl. authoring) | | | |
| *MNLI* | 22,445 | | 15,040 | |
| ours | 14,859 | ↓ 33.8% | 9,814 | ↓ 34.8% |

Table 2: Number of verified pairs in three-round validation phases. %cum-cons is the percentage of cumulative pairs with consensus gold label after completing each round. We show the efficiency of our proposed three-round validation compared to MNLI-style

In contrast to most previous NLI studies that only use a single sentence as the premise, we use premise text consists of a varying number of sentences, i.e., single-sentence (SINGLE-SENT), double-sentence (DOUBLE-SENTS), and multiple sentences (PARAGRAPH).[2]

### 3.2 Annotation Protocol

To collect NLI data for Indonesian, we follow the data collection protocol used in SNLI, MNLI, and OCNLI. It consists of two phases, i.e., hypothesis writing and label validation.

The annotation process involves two groups of annotators. We involve 27 Computer Science students as volunteers in the data collection project. All of them are native Indonesian speakers and were taking NLP classes. Henceforth, we refer to them as the **lay annotators**. The other group of annotators, which we call as **expert annotators** are five co-authors of this paper, who are also Indonesian native speakers and have at least seven years of experience in NLP.

**Writing Phase** In this phase, each annotator is assigned with 100-120 premises. For each premise, annotators are asked to write six hypothesis sentences, two for each semantic label (*entailment*, *contradiction*, and *neutral*). This strategy is similar to the MULTI strategy introduced in the OCNLI data collection protocol.[3] We provide instruction used in the writing phase in Appendix D.

For hypothesis writing involving expert annotators, we further ask the annotators to tag linguistic

[2]We limit PARAGRAPH to have a maximum 200-word length so that the current pre-trained language model for Indonesian can process it.

[3]In OCNLI, three hypothesis sentences per label are created for each premise, resulting in a total of nine hypotheses.

| | SNLI* | MNLI* | XNLI$_{EN}$* | OCNLI§ | IndoNLI | |
|---|---|---|---|---|---|---|
| | | | | | Lay | Expert |
| #pairs in total | 570,152 | 432,702 | 7,500 | 56,525 | 14,728 | 3,008 |
| #pairs validated | 56,941 | 40,000 | 7,500 | 9,913 | 4,489 | 3,008 |
| % validated per total | 10.0% | 9.2% | 100.0% | 17.5% | 30.5% | 100.0% |
| % pairs with gold label | 98.0% | 98.2% | 93.0% | 98.6% | 98.0% | 99.2% |
| individual label = gold label | 89.0% | 88.7% | n/a | 87.5% | 88.8% | 91.2% |
| individual label = author's label | 85.8% | 85.2% | n/a | 80.8% | 86.2% | 89.0% |
| gold label = author's label | 91.2% | 92.6% | n/a | 89.3% | 90.6% | 94.0% |
| gold label ≠ author's label | 6.8% | 5.6% | n/a | 9.3% | 7.4% | 5.2% |
| no gold label (no 3 labels match) | 2.0% | 1.8% | n/a | 1.4% | 2.0% | 0.8% |

Table 3: IndoNLI data labeling agreement, compared to other NLI dataset. *The number for SNLI, MNLI, English subset of dev and test split of XNLI (XNLI$_{EN}$) are copied from original papers (Bowman et al., 2015; Williams et al., 2018; Conneau et al., 2018). §For OCNLI, we recalculate the aggregate number from the original paper (Hu et al., 2020) that provided detail for 4 different protocols

phenomena required to perform inference on each sentence pair. The linguistic phenomena include lexical change, syntactic structure, and semantic reasoning.We also ask the expert annotators to ensure that the generated premise-hypothesis pairs are reasonably distributed among different linguistic phenomena. Enforcing balanced distribution among all phenomena is challenging because not all phenomena can be applied to the given premise text. We expect this strategy to help us create non-trivial examples covering various linguistic phenomena in the Indonesian language.

**Validation Phase** We perform label verification for ~30% and 100% pairs of lay-authored and expert-authored data, respectively. Our validation process is done through three rounds. In the first round, each pair is relabeled by two other independent annotators. If the label determined by those two annotators is the same as the initial label given by the annotator in the writing phase (author), we assign it as the *gold label*. Otherwise, we move the sentence pair to the second round in which another different annotator provides the label to the data. If any label was chosen by three of the four annotators (i.e., author, two annotators in the first round, and another annotator in the second round), it is assigned as the gold label. If there is no consensus, we proceed to the last round to collect another label from the fourth annotator.

In the MNLI data collection protocol, the goal of the validation phase is to obtain a three-vote consensus from the original label by the author and the labels given by four other annotators. Therefore, the annotation needed under the MNLI protocol is $5N$ for $N$ pairs of data. For INDONLI, our three-round annotation process can reduce the number of required annotations to $3N + X + Y$, where $X$ and $Y$ are the numbers of data for the second and third annotation rounds, respectively.

Table 2 shows that approximately 15K annotations are required to label and validate 3K data in EXPERT data if we use MNLI-style validation process, while this number can be reduced into less than 10K annotations (34% more efficient) using our three-round annotation process.[4] Our proposed strategy requires less annotation cost, which is worthwhile for the NLP research community with a limited budget.

Table 3 summarizes our final data, along with a comparison to SNLI, MNLI, XNLI, and OCNLI. Our results are on par with the number reported in SNLI / MNLI and better than OCNLI. About 98% of the validated pairs have the gold label, suggesting that our dataset is of high quality in general. The annotator agreement for EXPERT data is higher than LAY data, suggesting that the first is less ambiguous than the latter.

### 3.3 The Resulting Corpus

After filtering out premise-hypothesis pairs with no gold labels (no consensus), we ended up with 17,712 annotated sentence pairs. All expert-annotated pairs possessing gold labels are used as a test set. The lay-annotated pairs are split into development and test sets, such that there is no overlapping premise text between both sets. In the end, we have two separate test sets: Test$_{EXPERT}$and Test$_{LAY}$. Sentence pairs that are not included in

[4]If we omit the number of annotations in the writing phase and only consider the number of annotations in the validation phase, the efficiency rate is even higher (> 40%).

| | Train | Dev | $\text{Test}_{\text{LAY}}$ | $\text{Test}_{\text{EXPERT}}$ |
|---|---|---|---|---|
| #entailment | 3476 | 807 | 808 | 1041 |
| #contradiction | 3439 | 749 | 764 | 999 |
| #neutral | 3415 | 641 | 629 | 944 |
| premise len | $21.0_{(14.0)}$ | $19.9_{(10.9)}$ | $20.4_{(11.6)}$ | $31.1_{(18.9)}$ |
| hypothesis len | $7.6_{(2.9)}$ | $7.7_{(2.8)}$ | $7.7_{(3.1)}$ | $9.3_{(4.2)}$ |
| #SINGLE-SENT | 8368 | 1784 | 1836 | 1534 |
| #DOUBLE-SENTS | 1442 | 336 | 282 | 1043 |
| #PARAGRAPH | 520 | 77 | 83 | 407 |

Table 4: IndoNLI corpus statistics. Length is calculated at token-level, with a simple space-delimited tokens. Numbers in the bracket shows the standard deviation.

| | Jaccard | LCS | New token rate |
|---|---|---|---|
| $\text{Test}_{\text{LAY}}$ | | | |
| Entailment | 31.8 | 71.4 | 16.7 |
| Contradiction | 28.6 | 66.7 | 25.0 |
| Neutral | 21.1 | 54.5 | 37.5 |
| $\text{Test}_{\text{EXPERT}}$ | | | |
| Entailment | 21.1 | 60.0 | 30.0 |
| Contradiction | 20.8 | 62.5 | 28.6 |
| Neutral | 15.1 | 44.4 | 46.2 |

Table 5: Word overlap between premise and hypothesis in $\text{Test}_{\text{LAY}}$ vs. $\text{Test}_{\text{EXPERT}}$.

the validation phase and the lay-annotated pairs without a gold label are used for the training set.[5] The number of expert-annotated pairs missing gold labels is extremely small. We excluded them in the distributed corpus. INDONLI data characteristics is described in the Appendix A (Bender and Friedman, 2018)

The resulting corpus statistic is presented in Table 4. We observe that the three semantic labels have a relatively balanced distribution. On average, lay-annotated data seems to have a shorter premise and hypothesis length than expert-annotated data. In both LAY and EXPERT data, single-sentence premises (SINGLE-SENT) is the most dominant, followed by DOUBLE-SENTS and PARAGRAPH.

**Word Overlap Analysis** McCoy et al. (2019) show that NLI models might utilize lexical overlap between premise and hypothesis as a signal for the correct NLI label. To measure this in our data, we use the Jaccard index to measure *unordered* word overlap and the longest common subsequence, LCS, to measure *ordered* word overlap. In addition, we also measure the new token rate (i.e., the percentage of hypothesis tokens not present in the premise) as a proxy to measure token diversity in the hypothesis. Table 5 shows our results. Regarding the Jaccard index, $\text{Test}_{\text{EXPERT}}$ has an overall lower similarity than $\text{Test}_{\text{LAY}}$ and the two have higher similarity for pairs with entailment labels than the other labels. $\text{Test}_{\text{EXPERT}}$ also has a lower LCS similarity score than $\text{Test}_{\text{LAY}}$, suggesting that the expert annotators use different wording or sentence structure in the hypothesis. In terms of the new token rate, we find that $\text{Test}_{\text{EXPERT}}$has a higher rate than $\text{Test}_{\text{LAY}}$. This indicates that, in general, expert annotators use more diverse tokens than lay annotators when generating hypotheses.

## 4 Experiments

We experiment with several neural network-based models to evaluate the difficulty of our corpus. As our baseline, we use a continuous bag of words (CBoW) model, initialized with Indonesian fastText embeddings (Bojanowski et al., 2017). The others are Transformers-based models: multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and two pre-trained models for Indonesian, IndoBERT and IndoBERT-lite (Wilie et al., 2020). The first two are multilingual models which are trained on multilingual data, which includes Indonesian text. IndoBERT uses BERT large architecture with 335.2M parameters, while IndoBERT-lite uses ALBERT (Lan et al., 2020) architecture with fewer number of parameters (11.7M).[6]

**Training and Optimization** We use Adam (Kingma and Ba, 2015) optimizer for training all models. For the experiment with the CBoW model, we use a learning rate $1 \times 10^{-4}$, batch size of 8, and dropout rate 0.2. For experiments with Transformers models, we perform hyperparameter sweep on the learning rate $\in \{1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-6}\}$ and use batch size of 16. Each model is trained for 10 epochs with early stopping, with three random restarts.

For all of our experiments, we use the `jiant` toolkit (Phang et al., 2020), which is based on Pytorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020). We use NVIDIA V100 Tensor Core GPUs for our experiments. More details regarding training time can be found in the Appendix B.

[5] We use the initial label given by the author.

[6] We use indobert-large-p2 and indobert-lite-base-p2 models for our experiments.

| | Dev | $\text{Test}_{\text{LAY}}$ | $\text{Test}_{\text{EXPERT}}$ |
|---|---|---|---|
| Human | - | *85.1*$_{(0.0)}$ | *89.1*$_{(0.0)}$ |
| Majority | 35.7$_{(0.0)}$ | 36.7$_{(0.0)}$ | 34.9$_{(0.0)}$ |
| CBoW | 54.7$_{(0.2)}$ | 50.1$_{(0.5)}$ | 36.9$_{(0.3)}$ |
| $\text{IndoBert}_{\text{lite}}$ | 76.2$_{(0.5)}$ | 74.1$_{(0.2)}$ | 58.9$_{(0.9)}$ |
| $\text{IndoBert}_{\text{large}}$ | 78.7$_{(0.4)}$ | 77.1$_{(0.5)}$ | 61.5$_{(2.2)}$ |
| mBERT | 76.2$_{(0.8)}$ | 72.5$_{(0.4)}$ | 57.3$_{(0.8)}$ |
| XLM-R | **85.7**$_{(0.4)}$ | **82.3**$_{(0.3)}$ | **70.3**$_{(1.0)}$ |

Table 6: Average model accuracy on the development and test sets over three random restarts. Numbers in the bracket shows the standard deviation. See Appendix C for the detail about human baseline.

| | $\text{Test}_{\text{LAY}}$ | $\text{Test}_{\text{EXPERT}}$ |
|---|---|---|
| *By Label* | | |
| Entailment | 81.4 | 56.6 |
| Contradiction | 82.7 | 63.3 |
| Neutral | 83.1 | 90.1 |
| *By Premise Sentence Type* | | |
| SINGLE-SENT | 82.3 | 70.9 |
| DOUBLE-SENTS | 83.3 | 69.0 |
| PARAGRAPH | 79.5 | 65.3 |

Table 7: Performance comparison of lay and expert data based on the NLI label and the premise type.

| | Dev | $\text{Test}_{\text{LAY}}$ | $\text{Test}_{\text{EXPERT}}$ |
|---|---|---|---|
| $\text{IndoBERT}_{\text{lite}}$ | 57.6$_{(0.4)}$ | 56.7$_{(0.5)}$ | 45.9$_{(0.4)}$ |
| $\text{IndoBERT}_{\text{large}}$ | 60.3$_{(0.9)}$ | 59.5$_{(1.2)}$ | 45.7$_{(1.7)}$ |
| mBERT | 57.5$_{(0.8)}$ | 56.9$_{(1.1)}$ | 44.5$_{(1.0)}$ |
| XLM-R | **60.5**$_{(0.5)}$ | **59.8**$_{(1.0)}$ | **46.0**$_{(0.8)}$ |

Table 8: Hypothesis-only baseline results.

## 5 Results and Analysis

### 5.1 Model Performance

Table 6 reports the performance of all models on the INDONLI dataset, along with human performance on both test sets (Nangia and Bowman, 2019). CBoW model gives moderate improvements over the majority baseline. However, as expected, its performance is still below the Transformers model performance. We observe that $\text{IndoBERT}_{\text{lite}}$ obtains comparable or better performance than mBERT. $\text{IndoBERT}_{\text{large}}$ has better performance than $\text{IndoBERT}_{\text{lite}}$, but worse than XLM-R. This is interesting since one might expect that the Transformers model trained only on Indonesian text will give better performance than a multilingual Transformers model. One possible explanation is because the size of Indonesian pretraining data in XLM-R is much larger than the one used in $\text{IndoBERT}_{\text{large}}$ (180GB vs. 23GB uncompressed). This finding is also in line with IndoNLU benchmark results (Wilie et al., 2020), where XLM-R outperforms $\text{IndoBERT}_{\text{large}}$ on several tasks.

In terms of difficulty, it is evident that $\text{Test}_{\text{EXPERT}}$ is more challenging than $\text{Test}_{\text{LAY}}$ as there is a large margin of performance (up to 16%) between $\text{Test}_{\text{EXPERT}}$ and $\text{Test}_{\text{LAY}}$ across all models. We also see larger human-model gap in $\text{Test}_{\text{EXPERT}}$ (18.8%) compared to $\text{Test}_{\text{LAY}}$(2.8%). This suggests that INDONLI is relatively challenging for all the models, as there is still room for improvements for $\text{Test}_{\text{LAY}}$ and even more for $\text{Test}_{\text{EXPERT}}$.

**Analysis by Labels and Premise Type** We compare the performance based on the NLI labels and premise type between $\text{Test}_{\text{LAY}}$ and $\text{Test}_{\text{EXPERT}}$ (Table 7). We observe that the accuracy across labels is similar on the lay data, while on $\text{Test}_{\text{EXPERT}}$, the performance between labels is substantially different. Overall, the neutral label is relatively easier to predict than other labels for both $\text{Test}_{\text{LAY}}$ and $\text{Test}_{\text{EXPERT}}$. Contradiction and entailment labels for $\text{Test}_{\text{EXPERT}}$ are considerably more difficult than $\text{Test}_{\text{LAY}}$. For the performance based on the premise sentence type, there is no substantial accuracy difference between SINGLE-SENT and DOUBLE-SENTS for both $\text{Test}_{\text{EXPERT}}$ and $\text{Test}_{\text{LAY}}$. When moving to multiple-sentence premise type (PARAGRAPH), we observe a large drop in performance for both $\text{Test}_{\text{EXPERT}}$ and $\text{Test}_{\text{LAY}}$.

### 5.2 Annotation Artifacts

**Hypothesis-only Models** Poliak et al. (2018b) propose a hypothesis-only model as a baseline when training an NLI model to investigate statistical patterns in the hypothesis that may reveal the actual label to the model. Table 8 shows results when we only use hypothesis as input to our models. On $\text{Test}_{\text{LAY}}$ split, our best performing model achieves ~60%, slightly lower than other NLI datasets (MNLI ~ 62%; SNLI ~ 69%, and OCNLI ~66%). We see much lower performance on the $\text{Test}_{\text{EXPERT}}$ split, with performance reduction up to 14%. This result indicates that our protocol for collecting $\text{Test}_{\text{EXPERT}}$ effectively reduces annotation artifacts that $\text{Test}_{\text{LAY}}$ has. However, since we do not use expert-written examples for training, $\text{Test}_{\text{EXPERT}}$ may have different artifacts that our models do not learn.

| Word | | Label | PMI | Count |
|---|---|---|---|---|
| **LAY** | | | | |
| salah | *wrong* | E | 0.72 | 96/160 |
| sekitar | *around* | E | 0.72 | 40/60 |
| suatu | *something* | E | 0.58 | 32/53 |
| bukan | *no* | C | 1.28 | 279/324 |
| tidak | *no* | C | 1.24 | 2319/2905 |
| apapun | *anything* | C | 1.20 | 67/70 |
| selain | *aside from* | N | 1.08 | 66/80 |
| juga | *also* | N | 1.05 | 109/146 |
| banyak | *a lot* | N | 0.94 | 291/452 |
| **EXPERT** | | | | |
| beberapa | *some* | E | 0.65 | 40/65 |
| dapat | *can* | E | 0.50 | 44/84 |
| ajaran | *doctrine* | E | 0.48 | 12/17 |
| tidak | *no* | C | 0.84 | 205/329 |
| kurang | *less* | C | 0.50 | 23/40 |
| didirikan | *established* | C | 0.49 | 14/21 |
| banyak | *a lot* | N | 0.69 | 54/90 |
| ia | *he/she* | N | 0.67 | 32/50 |
| juga | *also* | N | 0.63 | 37/62 |

Table 9: Top 3 PMI values between words and label on lay and expert data. **E**: Entailment, **C**: Contradiction, **N**: Neutral.

| | Dev | $\text{Test}_{\text{LAY}}$ | $\text{Test}_{\text{EXPERT}}$ |
|---|---|---|---|
| *Zero-shot* | $80.8_{(0.0)}$ | $78.3_{(0.0)}$ | $73.8_{(0.0)}$ |
| *Translate-train* | $82.8_{(1.1)}$ | $80.4_{(0.9)}$ | $\mathbf{75.7}_{(0.8)}$ |
| *Translate-train-s* | $79.3_{(0.9)}$ | $76.7_{(0.5)}$ | $71.1_{(0.2)}$ |
| INDONLI | $\mathbf{85.7}_{(0.4)}$ | $\mathbf{82.3}_{(0.3)}$ | $70.3_{(1.0)}$ |

Table 10: Comparison with *zero-shot* and *translate-train* approaches.

**PMI Analysis** We compute Pointwise Mutual Information (PMI) to see the discriminative words for each NLI label (Table 9). Manual analysis suggests that some words are actually part of multi-word expression (Suhardijanto et al., 2020). For example, the word *salah* is actually part of the expression *salah satu* which means *one of*. In general, we observe that the PMI values in lay data are relatively higher than expert data, indicating that the expert data has better quality and is more challenging. For contradiction label, it is dominated by negation words (e.g., *bukan*, *tidak* ). However, for expert data, only one negation word presents in the top 3 words, while for lay data, all top three words are negation words. This suggests that our annotation protocol in constructing expert data is effective in reducing these particular annotation artifacts.

## 6 Cross-Lingual Transfer Performance

Prior work (Conneau et al., 2018; Budur et al., 2020) has demonstrated the effectiveness of cross-lingual transfer when training data in the target language is not available. To evaluate the difficulty of our test sets in this setting, we experiment with two cross-lingual transfer approaches, zero-shot learning, and translate-train. In this experiment, we only use XLM-R as it obtains the best performance in our NLI evaluation.

In the *zero-shot learning*, we employ an XLM-R model trained using a concatenation of MNLI training set and XNLI validation set, which covers 15 languages in total.[7] In the *translate-train* setting, we machine-translate MNLI training and validation sets into Indonesian and fine-tune the pre-trained XLM-R on the translated data. Our English to Indonesian machine translation system uses the standard Transformer architecture (Vaswani et al., 2017) with 6 layers of encoders and decoders. Following Guntara et al. (2020), we train our translation model on a total of 13M pairs of multi-domain corpus from news articles, religious texts, speech transcripts, Wikipedia corpus, and back-translation data from OSCAR (Ortiz Suárez et al., 2020). We use Marian (Junczys-Dowmunt et al., 2018) toolkit to train our translation model.

Translate-train outperforms a model trained on our training data (INDONLI) on $\text{Test}_{\text{EXPERT}}$. Translate-train obtains the best performance with 5.4 points over the INDONLI model. We further investigate if the performance gap comes from the larger training data used for training our translate-train model. We train another model (*translate-train-s*), using a subset of translated training data such that the size is comparable to INDONLI training data. We find that the model performance using the same training data is also higher, although the gap is smaller (0.8 points). Since we do not include expert data in the INDONLI training data, this result indicates that the translated training data might contain more examples with similar characteristics with examples in $\text{Test}_{\text{EXPERT}}$ than INDONLI training data. Overall, we observe that the best performance on our $\text{Test}_{\text{EXPERT}}$ is still relatively low (e.g., 75.7 compared to the best performance in OCNLI, 78.2), indicating that the test set is still challenging.

[7] We use a pre-trained model distributed by HuggingFace: https://huggingface.co/joeddav/xlm-roberta-large-xnli.

| Inference tags | | #pairs (E:C:N) | | | | Accuracy (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | IndoBERT$_{\text{large}}$ | XLM-R | *translate-train* |
| Morphological derivation | MORPH | 96 | (47 | 31 | 18) | $69.4_{(3.2)}$ | $79.2_{(1.8)}$ | $84.4_{(1.0)}$ |
| Syntactic structure reordering | STRUCT | 100 | (53 | 36 | 11) | $60.7_{(2.5)}$ | $73.3_{(1.5)}$ | $79.0_{(2.6)}$ |
| Lexical subsequence | LSUB | 99 | (48 | 42 | 9) | $66.7_{(7.1)}$ | $74.4_{(0.6)}$ | $85.2_{(3.1)}$ |
| Negation | NEG | 75 | (11 | 55 | 9) | $71.1_{(0.8)}$ | $71.6_{(0.8)}$ | $79.1_{(3.4)}$ |
| Coordinating conjunction | COORD | 38 | (14 | 13 | 11) | $69.3_{(8.5)}$ | $80.7_{(1.5)}$ | $85.1_{(4.0)}$ |
| Logical quantification | QUANT | 59 | (18 | 20 | 21) | $68.9_{(2.6)}$ | $75.7_{(1.0)}$ | $75.1_{(3.5)}$ |
| Numerical & math reasoning | NUM | 120 | (40 | 43 | 37) | $45.0_{(2.2)}$ | $58.3_{(2.2)}$ | $65.0_{(0.0)}$ |
| Comparative & superlative | COMP | 51 | (12 | 15 | 24) | $59.5_{(4.1)}$ | $64.1_{(3.0)}$ | $59.5_{(7.9)}$ |
| Lexical semantics | LEXSEM | 166 | (68 | 73 | 25) | $58.4_{(4.2)}$ | $71.3_{(0.9)}$ | $76.9_{(3.0)}$ |
| Idiomatic expression | IDIOM | 28 | (12 | 3 | 13) | $59.5_{(5.5)}$ | $69.0_{(5.5)}$ | $78.6_{(3.6)}$ |
| Anaphora & coreference | COREF | 70 | (29 | 23 | 18) | $64.8_{(5.0)}$ | $74.3_{(1.4)}$ | $74.8_{(2.2)}$ |
| Spatial reasoning | SPAT | 37 | (11 | 8 | 18) | $45.9_{(2.7)}$ | $57.7_{(1.6)}$ | $68.5_{(5.6)}$ |
| Temporal expression & reasoning | TEMP | 68 | (15 | 20 | 33) | $56.4_{(8.9)}$ | $68.1_{(2.2)}$ | $69.6_{(3.7)}$ |
| Common-sense reasoning | CS | 105 | (42 | 18 | 45) | $55.6_{(3.1)}$ | $63.8_{(1.6)}$ | $70.2_{(1.5)}$ |
| World knowledge | WORLD | 70 | (16 | 20 | 34) | $67.1_{(3.8)}$ | $73.3_{(0.8)}$ | $69.0_{(1.6)}$ |

Table 11: Inference tags examined in INDONLI diagnostic set and model performance measured on pairs annotated with tag

## 7 Linguistic Phenomena in Test$_{\text{EXPERT}}$

To investigate natural language challenges in INDONLI data, we perform an in-depth analysis of linguistic phenomena and task accuracy breakdown (Linzen, 2020) on our test-bed. Specifically, we examine the distribution of inference categories in Test$_{\text{EXPERT}}$ data and investigate which category the models succeed or fail.

We curate a subset of 650 examples from Test$_{\text{EXPERT}}$ as the diagnostic dataset.[8] To annotate the diagnostic set with linguistics phenomena, we ask one expert annotator (who is not the example's author) to review the inference categories tagged by the expert-author when creating premise-hypotheses pairs (Williams et al., 2020). The annotation is multi-label, in which a premise-hypotheses pair can correspond to more than one natural language phenomenon.

Our annotation scheme incorporates 15 types of inference categorization. They include a variety of linguistic and logical phenomena and may require knowledge beyond text. The definition for inference tags examined in diagnostic set is stated in the Table 13 in the Appendix F. We provide the tag distribution in diagnostic set and also report the performance of IndoBERT$_{\text{large}}$, XLM-R, and *translate train* models on the curated examples, as shown in Table 11.

We see that many premise-hypothesis pairs in INDONLI diagnostic set apply lexical semantics (e.g., synonyms, antonyms, and hypernyms-hyponyms) or require common-sense knowledge (i.e., a basic understanding of physical and social dynamics) to make an inference. Pairs with NUM tag also occur with high frequency in our data, whereas the challenges of idiomatic expression is less prevalent. Few phenomena are evenly distributed among labels, e.g., NUM, QUANT, and COORD. On the other hand, there is only a small proportion of pairs in LSUB and STRUCT categories which have *neutral* labels. Many examples of NEG unsurprisingly have *contradiction* label.

Our analysis on diagnostic set shows that the models can handle examples tagged with morpho-syntactic categories or boolean logic well. COORD and MORPH are among 2 tags with highest performance for XLM-R. The translate-train model achieves 85% on LSUB, indicating that our data is considerably robust with respect to syntactic heuristics, such as lexical overlap and subsequence (McCoy et al., 2019). On the other hand, all models also have decent accuracy on inference pairs with negation; the performance remains stable in three different models (more than 70%).

In contrast, the hardest overall categories appear to be NUM and COMP, indicating that models struggle with arithmetic computation, reasoning about quantities, and dealing with comparisons (Ravichander et al., 2019; Roy et al., 2015). In addition, models find it difficult to reason about temporal and spatial properties, as the accuracy for examples annotated with TEMP and SPAT are also inadequate. Commonsense reasoning is shown as another challenge for NLI model, suggesting that there is still much room for improvement for models to learn tacit knowledge from the text.

[8] We use the same examples to evaluate human baseline

## 8 Conclusion

We present INDONLI, the first human-elicited NLI data for Indonesian. The dataset is authored and annotated by crowd (*lay data*) and expert annotators (*expert data*). INDONLI includes nearly 18K sentence pairs, makes it the largest Indonesian NLI dataset to date. We evaluate state-of-the-art NLI models on INDONLI, and find that our dataset, especially the one created by *expert* is challenging as there is still a substantial human-model gap on $\text{Test}_{\text{EXPERT}}$. The expert data contains more *diverse hypotheses* and *less annotation artifacts* makes it ideal for testing models beyond its normal capacity (*stress test*). Furthermore, our qualitative analysis shows that the best model struggles in handling linguistic phenomena, particularly in numerical reasoning and comparatives and superlatives. We expect this dataset can contribute to facilitating further progress in Indonesian NLP research.

## 9 Ethical Considerations

INDONLI is created using premise sentences taken from Wikipedia, news, and web domains. These data sources may contain harmful stereotypes, and thus models trained on this dataset have the potential to reinforce those stereotypes. We argue that additional measurement on the potential harms introduced by these stereotypes is needed before using this dataset to train and deploy models for real-world applications.

## Acknowledgments

We would like to thank Kerenza Doxolodeo, Theresia Veronika Rampisela, and Ajmal Kurnia for their contribution in preparing unannotated data set and assisting the annotation process. We also thank the student annotators, without whom this work would not have been possible.

RM's work on this project was financially supported by a grant from Program Kompetisi Kampus Merdeka (PKKM) 2021, Faculty of Computer Science, Universitas Indonesia.

CV's work on this project at New York University was financially supported by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program) and Samsung Research (under the project *Improving Deep Learning using Latent Structure*) and benefitted from in-kind support by the NYU High-Performance Computing Center. This material is based upon work supported by the National Science Foundation under Grant No. 1922658. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

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

Hossein Amirkhani, Mohammad Azari Jafari, Azadeh Amirak, Zohreh Pourjafari, Jahromi Soroush Faridan, and Zeinab Kouhkan. 2020. Farstail: A persian natural language inference. *arXiv preprint arXiv:2009.08820*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2010. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In

*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Gimenez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

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

Tri Wahyu Guntara, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43, Marseille, France. European Language Resources Association.

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.

Rani Aulia Hidayat, Isnaini Nurul Khasanah, Wava Carissa Putri, and Rahmad Mahendra. 2021. Feature-rich classifiers for recognizing textual entailment in indonesian. *Procedia Computer Science*, 189:148–155. AI in Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

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

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16, Marseille, France. European Language Resources Association.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pages 215–220. IEEE.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020*.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Lrec*, pages 216–223.

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.

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

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

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 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 337–340, 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.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*

*2020, New York, NY, USA, February 7-12, 2020*, pages 8713–8721. AAAI Press.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online. Association for Computational Linguistics.

Ken Nabila Setya and Rahmad Mahendra. 2018. Semi-supervised textual entailment on indonesian wikipedia data. In *2018 International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Totok Suhardijanto, Rahmad Mahendra, Zahroh Nuriah, and Adi Budiwiyanto. 2020. The framework of multiword expression in Indonesian language. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 582–588, Hanoi, Vietnam. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

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. 2020. 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. 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.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. Anlizing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*.

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. 2020. 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. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc,

Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

# A Data Statement

## A.1 Curation Rationale

INDONLI is the first human-elicited natural language inference (NLI) dataset for Indonesian. It is created to foster research in Indonesian NLP, especially for NLI.

The dataset consists of 17,712 premise-hypothesis pairs and is divided into training, development, and test splits. We curate two separate test sets, $\text{Test}_{\text{LAY}}$, which is authored and annotated by 27 students, and $\text{Test}_{\text{EXPERT}}$, which is authored and annotated by experts. The students were offered the compensation whose the rate is similar to the wage for a research assistant in Faculty of Computer Science, Universitas Indonesia. All authors and annotators are Indonesian native speakers.

## A.2 Language Variety

The premise sentences used to create INDONLI are taken from three sources: Wikipedia, news, and web domain. For the news text, we use premise sentences from the Indonesian PUD[9] and GSD[10] treebanks provided by the Universal Dependencies 2.5 (Zeman et al., 2019) and IndoSum dataset (Kurniawan and Louvan, 2018). For the web domain, we collect premise sentences from blogs and institutional websites (e.g., government, university, and school). Our manual analysis shows that most of the sentences are written in standard written Indonesian.

## A.3 Speaker Demographic

All INDONLI authors are Indonesian native speakers. Lay authors are undergraduate students who have taken NLP class, while expert authors consist of 2 Ph.D. students and 3 researchers with at least 7 years experience in NLP. Besides this information, we do not collect any other demographic information of the authors.

## A.4 Annotator Demographic

The authors of INDONLI are also the annotators.

## A.5 Speech Situation

Each hypothesis in INDONLI is written based on premise sentence(s) taken from Wikipedia, news, or web articles.

[9] https://github.com/UniversalDependencies/UD_Indonesian-PUD

[10] https://github.com/UniversalDependencies/UD_Indonesian-GSD

## A.6 Text Characteristics

The premise text in INDONLI can be categorized into three groups: single-sentence (SINGLE-SENT), double-sentence (DOUBLE-SENTS),) and multiple-sentence (PARAGRAPH).

## A.7 Recording Quality

N/A

## A.8 Other

N/A

## A.9 Provenance Appendix

N/A

# B Training Time

| Model | Training Data | |
|---|---|---|
| | IndoNLI (hrs) | Indo_XNLI (days) |
| mBERT | ±3 | ±1 |
| XLM-R | ±10 | ±4 |
| $\text{IndoBERT}_{\text{large}}$ | ±3 | ±1 |
| $\text{IndoBERT}_{\text{lite}}$ | ±6 | ±3 |

Table 12: Training time for a single run on each model on a particular training data.

# C Determining Human Baseline

We follow the procedure in Nangia and Bowman (2019) to measure human baselines performance on INDONLI. We hire 3 Indonesian native speakers who do not participate in the data collection process. We provide them with 15 examples of labeled premise-hypothesis pairs (5 pairs for each label). We also tailor a short prompt explaining the NLI task definition. After reviewing the examples and prompt, the annotators are then given a stratified sample of 450 and 650 examples from $\text{Test}_{\text{LAY}}$ and $\text{Test}_{\text{EXPERT}}$, respectively. The gold label for a total of 1,100 examples are concealed, and the annotators are asked to perform labeling experiment. We compute the majority label from them and compare that against the gold label in INDONLI test data to obtain accuracy. For pairs with no majority label, we use the most frequent label from INDONLI test data (*entailment*).

## D INDONLI Writing Instruction

In this task, given a premise text consisting of one or more sentences, you are asked to write six different hypothesis sentences, two for each label (*entailment, contradiction, and neutral*).

A premise-hypothesis pair is annotated with ***entailment*** label if it can be concluded that the hypothetical text is correct based on the information contained in the premise text. It is annotated with ***contradiction*** label is if it can be concluded that the hypothesis text is wrong based on the information contained in the premise text. Otherwise, the label is ***neutral***; in other words, based on the information contained in the premise text, the truth of the hypothesis text cannot be determined (not enough information).

Please make sure that each hypothesis sentence satisfies the following criteria:
- It consists of one sentence and not multiple sentences,
- It contains some keywords present in the premise text,
- It is grammatical according to Indonesian grammar.

Some strategies that you can apply when writing the hypothesis sentence including, but not limited to:

1. *Word deletion*
   Delete one or more words from the premise text.

2. *Word addition*
   Add one or more words to the premise text. For example, you can add adjectives, negation words, etc.

3. *Lexical change*
   Replace one or more words from premise text with their synonym, antonym, hypernym, or hyponym.

4. *Paraphrase*
   Write premise text with your own words.

5. *Structural change*
   Change the structure of the premise text. For example, you can change the active voice into passive voice or change the order of the sub-sentence in the premise text.

6. *Reasoning*
   Apply reasoning to the given premise text to write a hypothesis sentence, such that the reasoning skill is needed when deciding the correct entailment label. For example, you can use numerical reasoning or commonsense knowledge.

If you are given a premise text which consists of multiple sentences, you *should not* write a hypothesis sentence that is *identical* to one of the sentences in the premise text.

Figure 1: This is the instruction given to lay authors for writing hypothesis sentences.

## E INDONLI Validation Instruction

In this task, you will be given a set of sentence pairs. For each sentence pair:

1. Check if they are free of errors such as ungrammatical, incomplete, or have wrong punctuation.

2. Fix any error that is found in one or both sentences.

3. Pick the correct semantic label for the sentence pair.

Figure 2: This is the instruction given to annotators for labeling each sentence pair.

## F The Linguistic Phenomena Examined in IndoNLI

| Our Tag | Description | Similar Tag in Related Work |
|---|---|---|
| MORPH | Transforming the word form by applying morphological derivation. For example, a noun into verb (verbalization) or vice versa (nominalization). | Lexical:Verbalization (Bentivogli et al., 2010) |
| STRUCT | Reordering the structure of arguments in the premise, e.g. changing active into passive voices. | Alternations (Wang et al., 2018)<br>Syntactic (Joshi et al., 2020) |
| LSUB | Syntactic heuristics, i.e., word overlap and lexical subsequence. This phenomena is captured in the data whose the hypothesis is obtained by deleting or adding the words without changing the sentence structure | (McCoy et al., 2019)<br>Word overlap (Naik et al., 2018) |
| NEG | The use of negation words, e.g., *tidak*, *bukan* | (Hossain et al., 2020)<br>Negation (Kim et al., 2019)<br>Negation (Naik et al., 2018)<br>Negation (Richardson et al., 2020) |
| COORD | Logical inference about coordinating conjunctions (e.g., *dan*, *tetapi*, *atau*) that conjoin two or more conjuncts of varied syntactic categories (i.e., noun phrases, verb phrases, clauses) | (Saha et al., 2020), Coord. (Kim et al., 2019)<br>Coordinations (Williams et al., 2020)<br>Boolean (Joshi et al., 2020) |
| QUANT | Inferences from the natural language analogs of universal and existential quantification | Quantification (Wang et al., 2018)<br>Quantifier (Joshi et al., 2020) |
| COMP | Comparatives and superlatives expressing qualitative or quantitative differences between entities | Comp. & Super. (Williams et al., 2020)<br>Comp. (Kim et al., 2019)<br>Comparatives (Richardson et al., 2020) |
| LEXSEM | Inferences made possible by lexical information about synonyms, antonyms, and hypernym-hyponyms | Lexical Entailment (Wang et al., 2018)<br>Lexical (Bentivogli et al., 2010)<br>(Glockner et al., 2018)<br>Lexical (Joshi et al., 2020) |
| NUM | Numerical expression, such as cardinal and/or ordinal numbers, percentage and money. It also includes the mathematical reasoning, and counting of the entities. | (Ravichander et al., 2019)<br>Numeral Reasoning (Naik et al., 2018)<br>Counting (Kim et al., 2019)<br>Numeral (Williams et al., 2020) |
| COREF | Anaphora and coreferences between pronouns, proper names, (e.g. named entities) and noun phrases | Anaphora/Coreference (Wang et al., 2018)<br>Coreference (Joshi et al., 2020)<br>Reference (Williams et al., 2020) |
| IDIOM | Idiomatic expression. | Idioms (Williams et al., 2020) |
| SPAT | Spatial reasoning that involves places and spatial relations between entities; understanding the preposition of location and direction | Spatial (Kim et al., 2019)<br>Spatial (Joshi et al., 2020) |
| TEMP | Temporal reasoning that involves a common sense of time, for example, the duration an event lasts, the general time an activity is carried out and, the sequence of events | (Vashishtha et al., 2020) |
| CS | Commonsense knowledge that is expected to be possessed by most people, independent of cultural or educational background. This includes a basic understanding of physical and social dynamics, plausibility of events, and cause-effect relations | Common sense (Wang et al., 2018)<br>Plausibility (Williams et al., 2020) |
| WORLD | Reasoning that requires knowledge about named entities, knowledge about historical and cultural, current events; and domain-specific knowledge. | World knowledge (Wang et al., 2018)<br>Reasoning-Fact (Williams et al., 2020)<br>World (Joshi et al., 2020) |

Table 13: The list of linguistic tags in IndoNLI diagnostic set and reference to similar tags in previous work.

## G IndoNLI Diagnostic Set Examples

| Premise | Hypothesis | Label | Phenomena |
|---|---|---|---|
| Topan Molave telah menewaskan 36 orang dan menyebabkan 46 orang lainnya hilang di Vietnam.<br>*(Typhoon Molave has killed 36 people and left 46 others missing in Vietnam.)* | Angka kematian lebih sedikit ketimbang angka orang hilang.<br>*(The death rate is less than the number of missing persons.)* | E | COMP, NUM |
| Selain rumah yang rusak, area persawahan milik warga 5 hektare longsor dengan kedalaman 10 meter dan lebar 250 meter.<br>*(In addition to the damaged houses, a 5-hectare rice field owned by residents had a landslide with 10 meters depth and 250 meters width.)* | Sawah dan rumah warga mengalami kerusakan.<br>*(Rice fields and houses were damaged.)* | E | COORD, MORPH, STRUCT |
| Penyerang Juventus, Cristiano Ronaldo, merayakan gol ke gawang Cagliari, Minggu (22/11/2020).<br>*(Juventus striker, Cristiano Ronaldo, celebrates a goal against Cagliari, Sunday (11/22/2020).)* | Ronaldo melakukan perayaan atas gol yang ia buat.<br>*(Ronaldo celebrates the goal he made.)* | E | COREF, MORPH |
| Semua calon petahana Pilkada 2020 di 3 kabupaten di Yogyakarta dinyatakan kalah dalam rapat pleno penghitungan suara Komisi Pemilihan Umum (KPU).<br>*(All incumbent candidates for the 2020 Pilkada (regional election) in 3 districts in Yogyakarta were declared defeated in the plenary meeting of the General Election Commission (KPU) vote counting.)* | Ada calon petahana Pilkada 2020 di 3 kabupaten di Yogyakarta yang menang.<br>*(There is incumbent candidate for the 2020 Pilkada in 3 districts in Yogyakarta who won.)* | C | QUANT, LEXSEM |
| Kebijakan untuk membuka sekolah dikembalikan kepada pemerintah daerah dengan persetujuan orang-tua.<br>*(The policy to open schools is assigned to the local government with parental permission.)* | Pemerintah daerah dapat membuka sekolah tanpa persetujuan orang tua.<br>*(Local governments can open schools without parental permission.)* | C | NEG, STRUCT |
| Kota Gunungsitoli terletak di Pulau Nias dan berjarak sekitar 85 mil laut dari Kota Sibolga.<br>*(Gunungsitoli City is located on Nias Island and is about 85 nautical miles from Sibolga City.)* | Kota Sibolga terletak di Pulau Nias.<br>*(Sibolga City is located on Nias Island.)* | C | SPAT, STRUCT |
| Perlahan-lahan, keluarga Kim berusaha agar satu per satu anggota keluarga mereka dapat bekerja di keluarga Park, dengan saling merekomendasikan satu sama lain dan berbohong sebagai penyedia jasa profesional yang saling tidak kenal.<br>*(Gradually, the Kims try to get each of their family members to work for the Parks, recommending each other and lying as professional service providers who don't know each other.)* | Tipu daya keluarga Kim berujung di meja hijau.<br>*(The Kim family's trickery ended up at the court.)* | N | IDIOM, CS |
| Ismed Sofyan (lahir di Manyak Payed, Aceh Tamiang, 28 Agustus 1979; umur 41 tahun) adalah pemain Persija dan tim nasional Indonesia.<br>*(Ismed Sofyan (born in Manyak Payed, Aceh Tamiang, August 28, 1979; age 41) is a Persija player and the Indonesian national team.)* | Ismed Sofyan menghabiskan masa mudanya di Jakarta.<br>*(Ismed Sofyan spent his youth in Jakarta.)* | N | TEMP, WORLD |

Table 14: Annotated examples from the diagnostic set