IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP

Fajri Koto[1]   **Afshin Rahimi**[2]   **Jey Han Lau**[1]   **Timothy Baldwin**[1]

[1]The University of Melbourne
[2]The University of Queensland

ffajri@student.unimelb.edu.au, afshinrahimi@gmail.com
jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

Although the Indonesian language is spoken by almost 200 million people and the 10th most-spoken language in the world,[1] it is under-represented in NLP research. Previous work on Indonesian has been hampered by a lack of annotated datasets, a sparsity of language resources, and a lack of resource standardization. In this work, we release the INDOLEM dataset comprising seven tasks for the Indonesian language, spanning morpho-syntax, semantics, and discourse. We additionally release INDOBERT, a new pre-trained language model for Indonesian, and evaluate it over INDOLEM, in addition to benchmarking it against existing resources. Our experiments show that INDOBERT achieves state-of-the-art performance over most of the tasks in INDOLEM.

1 Introduction

Despite there being over 200M first-language speakers of the Indonesian language, the language is under-represented in NLP. We argue that there are three root causes: a lack of annotated datasets, a sparsity of language resources, and a lack of resource standardization. In English, on the other hand, there are ever-increasing numbers of datasets for different tasks (Hermann et al., 2015; Luong and Manning, 2016; Rajpurkar et al., 2018; Agirre et al., 2016), (pre-)trained models for language modelling and language understanding tasks (Devlin et al., 2019; Yang et al., 2019; Radford et al., 2019), and standardized tasks to benchmark research progress (Wang et al., 2019b; Wang et al., 2019a; Williams et al., 2018), all of which have contributed to rapid progress in the field in recent years.

We attempt to redress this situation for Indonesian, as follows. First, we introduce INDOLEM ("Indonesian Language Evaluation Montage"[2]), a comprehensive dataset encompassing seven NLP tasks and eight sub-datasets, five of which are based on previous work and three are novel to this work. As part of this, we standardize data splits and evaluation metrics, to enhance reproducibility and robust benchmarking. These tasks are intended to span a broad range of morpho-syntactic, semantic, and discourse analysis competencies for Indonesian, to be able to benchmark progress in Indonesian NLP. First, for morpho-syntax, we examine part-of-speech (POS) tagging (Dinakaramani et al., 2014), dependency parsing with two Universal Dependency (UD) datasets, and two named entity recognition (NER) tasks using public data. For semantics, we examine sentiment analysis and single-document summarization. For discourse, we create two Twitter-based document coherence tasks: Twitter response prediction (as a multiple-choice task), and Twitter document thread ordering.

Second, we develop and release INDOBERT, a monolingual pre-trained BERT language model for Indonesian (Devlin et al., 2019). This is one of the first monolingual BERT models for the Indonesian language, trained following the best practice in the field.[3]

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

[1]https://www.visualcapitalist.com/100-most-spoken-languages/

[2]Yes, guilty as charged, a slightly-forced backronym from *lem*, which is Indonesian for "glue", following the English benchmark naming trend (e.g. GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a)).

[3]Turns out we weren't the first to think to train a monolingual BERT model for Indonesian, or to name it IndoBERT, with (at least) two contemporaneous BERT models for Indonesian that are named "IndoBERT": Azhari and Lintang (2020) and Wilie et al. (2020).

Our contributions in this paper are: (1) we release INDOLEM, which is by far the most comprehensive NLP dataset for Indonesian, and intended to provide a benchmark to catalyze further NLP research on the language; (2) as part of INDOLEM, we develop two novel discourse tasks and datasets; and (3) we follow best practice in developing and releasing for general use INDOBERT, a BERT model for Indonesian, which we show to be superior to existing pre-trained models based on INDOLEM. The INDOLEM dataset, INDOBERT model, and all code associated with this paper can be accessed at: `https://indolem.github.io`.

2   Related Work

To comprehensively evaluate natural language understanding (NLU) methods for English, collections of tools and corpora such as GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) have been proposed. Generally, such collections aim to benchmark models across various NLP tasks covering a variety of corpus sizes, domains, and task formulations. GLUE comprises nine language understanding tasks built on existing public datasets, while SuperGLUE is a set of eight tasks that is not only diverse in task format but also includes low-resource settings. SuperGLUE is a more challenging framework, and BERT models trail human performance by 20 points at the time of writing.

In the cross-lingual setting, XGLUE (Liang et al., 2020) was introduced as a benchmark dataset that covers nearly 20 languages. Unlike GLUE, XGLUE includes language generation tasks such as question and headline generation. One of the largest cross-lingual corpora is dependency parsing provided by Universal Dependencies.[4] It has consistent annotation of 150 treebanks across 90 languages, constructed through an open collaboration involving many contributors. Recently, other cross-lingual benchmarks have been introduced, such as Hu et al. (2020) and Lewis et al. (2020). While these three cross-lingual benchmarks contain some resources/datasets for Indonesian, the coverage is low and data is limited.

Beyond the English and cross-lingual settings, ChineseGLUE[5] is a comprehensive NLU collection for Mandarin Chinese, covering eight different tasks. For the Vietnamese language, Nguyen and Nguyen (2020) gathered a dataset covering four tasks (NER, POS tagging, dependency parsing, and language inference), and empirically evaluated them against a monolingual BERT. Elsewhere, there are individual efforts to maintain a systematic catalogue of tasks and datasets, and state-of-the-art methods for each across multiple languages,[6] including one specifically for Indonesian.[7] However, there is no comprehensive dataset for evaluating NLU systems in the Indonesian language, a void which we seek to fill with INDOLEM.

3   INDOBERT

Transformers (Vaswani et al., 2017) have driven substantial progress in NLP research based on pre-trained models in the last few years. Although attention-based models are data- and GPU-hungry, the full attention mechanisms and parallelism offered by the transformer are highly compatible with the high levels of parallelism that GPU computation offers, and have been shown to be highly effective at capturing the syntax (Jawahar et al., 2019) and sentence semantics of text (Sun et al., 2019). In particular, transformer-based language models (Devlin et al., 2019; Radford et al., 2018; Conneau and Lample, 2019; Raffel et al., 2019) pre-trained on large volumes of text based on simple tasks such as masked word prediction and sentence ordering prediction, have quickly become ubiquitous in NLP and driven substantial empirical gains across tasks including NER (Devlin et al., 2019), POS tagging (Devlin et al., 2019), single document summarization (Liu and Lapata, 2019), syntactic parsing (Kitaev et al., 2019), and discourse analysis (Nie et al., 2019). However, this effect has been largely observed for high-resource languages such as English.

INDOBERT is a transformer-based model in the style of BERT (Devlin et al., 2019), but trained purely as a masked language model trained using the Huggingface[8] framework, following the default configura-

---

[4]`https://universaldependencies.org/`
[5]`https://github.com/ChineseGLUE/ChineseGLUE`
[6]`https://github.com/sebastianruder/NLP-progress`
[7]`https://github.com/kmkurn/id-nlp-resource`
[8]`https://huggingface.co/`

Data	#train	#dev	#test	5-Fold	Evaluation
Morpho-syntax/Sequence Labelling Tasks					
POS Tagging*	7,222	802	2,006	Yes	Accuracy
NER UI	1,530	170	425	No	micro-averaged F1
NER UGM	1,687	187	469	No	micro-averaged F1
UD-Indonesian GSD*	4,477	559	557	No	UAS, LAS
UD-Indonesian PUD (Corrected Version)	700	100	200	Yes	UAS, LAS
Semantic Tasks					
Sentiment Analysis	3,638	399	1,011	Yes	F1
IndoSum*	14,262	750	3,762	Yes	ROUGE
Coherency Tasks					
Next Tweet Prediction (NTP)	5,681	811	1,890	No	Accuracy
Tweet Ordering	5,327	760	1,521	Yes	Rank Corr

Table 1: Summary of datasets incorporated in INDOBERT. Datasets marked with '*' were already available with canonical splits.

tion for BERT-Base (uncased). It has 12 hidden layers each of 768d, 12 attention heads, and feed-forward hidden layers of 3,072d. We modify the Huggingface framework to read a separate text stream for different document blocks,[9] and set the training to use 512 tokens per batch. We train INDOBERT with 31,923-size Indonesian WordPiece vocabulary.

In total, we train INDOBERT over 220M words, aggregated from three main sources: (1) Indonesian Wikipedia (74M words); (2) news articles from Kompas,[10] Tempo[11] (Tala et al., 2003), and Liputan6[12] (55M words in total); and (3) an Indonesian Web Corpus (Medved and Suchomel, 2017) (90M words). After preprocessing the corpus into 512-token document blocks, we obtain 1,067,581 train instances and 13,985 development instances (without reduplication). In training, we use 4 Nvidia V100 GPUs (16GB each) with a batch size of 128, learning rate of 1e-4, the Adam optimizer, and a linear scheduler. We trained the model for 2.4M steps (180 epochs) for a total of 2 calendar months,[13] with the final perplexity over the development set being 3.97 (similar to English BERT-base).

4 INDOLEM: Tasks

In this section, we present an overview of INDOLEM, in terms of the NLP tasks and sub-datasets it includes. We group the tasks into three categories: morpho-syntax/sequence labelling, semantics, and discourse coherence. We summarize the sub-datasets include in INDOLEM in Table 1, in addition to detailing related work on the respective tasks.

4.1 Morpho-syntax and Sequence Labelling Tasks

Part-of-speech (POS) tagging. The first Indonesian POS tagging work was done over a 15K-token dataset. Pisceldo et al. (2009) defines 37 tags covering five main POS tags: *kata kerja* (verb), *kata sifat* (adjective), *kata keterangan* (adverb), *kata benda* (noun), and *kata tugas* (function words). They utilized news domain and partial data from the PanLocalisation project ("PANL10N"[14]). In total, "PANL10N" comprises 900K tokens, and was generated by machine-translating an English POS-tagged dataset and noisily projecting the POS tags from English to the Indonesian translations.

---

[9]The existing implementation merges all documents into one text stream

[10]https://kompas.com

[11]https://koran.tempo.co

[12]https://liputan6.com

[13]We checkpointed the model at 1M and 2M steps, and found that 2M steps yielded a lower perplexity over the dev set.

[14]http://www.panl10n.net/

To create a larger and more reliable corpus, Dinakaramani et al. (2014) published a manually-annotated corpus of 260K tokens (10K sentences). The text was sourced from the IDENTIC parallel corpus (Larasati, 2012), which was translated from data in the Penn Treebank corpus. The text is manually annotated with 23 tags based on Indonesian tag definition of Adriani et al. (2009). For INDOLEM, we use the Indonesian POS tagging dataset of Dinakaramani et al. (2014), and 5-fold partitioning of Kurniawan and Aji (2018).[15]

Named entity recognition (NER). Budi et al. (2005) was the first study on named entity recognition for Indonesian, where roughly 2,000 sentences from a news portal were annotated with three NE classes: `person`, `location`, and `organization`. In other work, Luthfi et al. (2014) utilized Wikipedia and DBPedia to automatically generate an NER corpus, and trained a model with Stanford CRF-NER (Finkel et al., 2005). Rachman et al. (2017) studied LSTM performance over 480 tweets with the same three named entity classes. None of these authors released the datasets used in the research.

There are two publicly-available Indonesian NER datasets. The first, NER UI, comprises 2,125 sentences obtained via an annotation assignment in an NLP course at the University of Indonesia in 2016 (Gultom and Wibowo, 2017). The corpus has the same three named entity classes as its predecessors (Budi et al., 2005). The second, NER UGM, comprises 2,343 sentences from news articles, and was constructed at the University of Gajah Mada (Fachri, 2014) based on five named entity classes: `person`, `organization`, `location`, `time`, and `quantity`.

Dependency parsing. Kamayani and Purwarianti (2011) and Green et al. (2012) pioneered dependency parsing for the Indonesian language. Kamayani and Purwarianti (2011) developed language-specific dependency labels based on 20 sentences, adapted from Stanford Dependencies (de Marneffe and Manning, 2016). Green et al. (2012) annotated 100 sentences of IDENTIC without dependency labels, and used an ensemble SVM model to build a parser. Later, Rahman et al. (2017) conducted a comparative evaluation over models trained using off-the-shelf tools such as MaltParser (Nivre et al., 2005) on 2,098 annotated sentences from the news domain. However, this corpus is not publicly available.

The Universal Dependencies (UD) project[16] has released two different Indonesian corpora of relatively small size: (1) 5,593 sentences of UD-Indo-GSD (McDonald et al., 2013);[17] and (2) 1,000 sentences of UD-Indo-PUD (Zeman et al., 2018).[18] Alfina et al. (2019) found that these corpora contain annotation errors and did not deal adequately with Indonesian morphology. They released a corrected version of UD-Indo-PUD by fixing annotations for reduplicated-words, clitics, compound words, and noun phrases.

We include two UD-based dependency parsing datasets in INDOLEM: (1) UD-Indo-GSD, and (2) the corrected version of UD-Indo-PUD. As our reference dependency parser model, we use the BiAffine dependency parser (Dozat and Manning, 2017), which has been shown to achieve strong performance for English.

4.2 Semantic Tasks

Sentiment analysis. There has been sentiment analysis for Indonesian domains/data sources including presidential elections (Ibrahim et al., 2015), stock prices (Cakra and Trisedya, 2015), Twitter (Koto and Rahmaningtyas, 2017), and movie reviews (Nurdiansyah et al., 2018). Most previous work, however, has used non-public and low-resource datasets.

We include in INDOLEM an Indonesian sentiment analysis dataset based on binary classification. In total, the data distribution is 3638/399/1011 sentences for train/development/test, respectively. The data was sourced from Twitter (Koto and Rahmaningtyas, 2017) and hotel reviews.[19] The hotel review data is annotated at the aspect level, where one review can have multiple polarities for different aspects. We

---

[15] We do not include POS data from the Universal Dependency project, as we found the data to contain many foreign borrowings (without any attempt to translate them into Indonesian), and some sentences to be poor translations (a point we return to in the context of error analysis of dependency parsing in Section 7).

[16] https://universaldependencies.org/

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

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

[19] https://github.com/annisanurulazhar/absa-playground/

Premise	Premise
• Ini kak Gracia sama kak Jerome saudaranya apa crushnya, serius tanya bang • Crush, karena awal mereka kenal karena sama2 dapet beasiswa mitsui	• Seriously ask, is Gracia Jerome's crush? or are they family? • His crush, they know each other when they got Mitsui scholarship
Possible next tweets:	Possible next tweets:
• selamat pagi min. Jika ingin bertanya terkait Latsar Cpns Kemdikbud lewat kontak mana ya?? • **Waw terimakasii, maaf aku followers baru.** • tahun ajaran barunya januari terus nnt aku masuk masuk ke sekolah di tanyain malah jadi tukang keong • lgi ap cantik?	• good morning admin. What is the contact for Latsar Cpns Kemdikbud?? • **Wow, thank you, sorry, I am a new follower.** • the new school academic year is January, on the first day I may get question about "tukang keong". • what are you doing beautiful?

Figure 1: Example for the next tweet prediction task. To the left is the original Indonesian version and to the right is an English translation. The tweet indicated in bold is the correct next tweet.

simply count the proportion of positive and negative polarity aspects, and label the sentence based on the majority sentiment. We discard a review if there is a tie in positive and negative aspects.

Summarization. From attention mechanisms (Rush et al., 2015; See et al., 2017) to pre-trained language models (Liu and Lapata, 2019; Zhang et al., 2019), recent summarization work on English in terms of both extractive and abstractive methods has relied on ever-larger datasets and data-hungry methods.

Indonesian (single document) text summarization research has inevitably focused predominantly on extractive methods, based on small datasets. Aristoteles et al. (2012) deployed a genetic algorithm over a 200-document summarization dataset, and Gunawan et al. (2017) performed unsupervised summarization over 3,075 news articles. As an attempt to create a standardized corpus, Koto (2016) released a 300-document chat summarization dataset, and Kurniawan and Louvan (2018) released the *IndoSum* 19K document–summary dataset. At the time we carried out this work,[20] *IndoSum* was the largest Indonesian summarization corpus in the news domain, manually constructed from CNN Indonesia[21] and Kumparan[22] documents. *IndoSum* is a single-document summarization dataset where each article has one abstractive summary. Kurniawan and Louvan (2018) released *IndoSum* together with the ORACLE — a set of extractive summaries generated automatically by maximizing ROUGE score between sentences of the article and its abstractive summary. We include *IndoSum* as the summarization dataset in INDOLEM, and evaluate the performance of extractive summarization in this paper.

4.3 Discourse Coherence Tasks

We also introduce two tasks that measure the ability of models to measure discourse coherence in Indonesian, based on message ordering in Twitter threads, namely: (1) next tweet prediction; and (2) message ordering. Utilizing tweets instead of edited text arguably makes the task harder and allows us to assess the robustness of models.

First, we use the standard twitter API filtered with the language parameter to harvest 9M Indonesian tweets from the period April–May 2020, covering the following topics: health, education, economy, and government. We discard threads that contain more than three self-replies, and threads containing similar tweets (usually from Twitter bots). Specifically, we discard a thread if 90% of the tweets are similar, as based on simple lexical overlap.[23] We gather threads that contain 3–5 tweets, and anonymize all mentions. This data is used as the basis for the two discourse coherence tasks.

Next tweet prediction. To evaluate model coherence, we design a next tweet prediction (NTP) task that is similar to the next sentence prediction (NSP) task used to train BERT (Devlin et al., 2019). In NTP, each instance consists of a Twitter thread (2–4 tweets) that we call the premise, and four possible options for the next tweet (see Figure 1 for an example), one of which is the actual response from the

---

[20]Noting that the soon-to-be-released Liputan6 dataset (Koto et al., to appear) will be substantially larger, but was not available when this research was carried out.

[21]https://www.cnnindonesia.com/

[22]https://kumparan.com/

[23]Two tweets are considered to be similar if they have a vocabulary overlap ≥80%.

original thread. In total, we construct 8,382 instances, where the distractors are obtained by randomly picking three tweets from the Twitter crawl. We ensure that there is no overlap between the next tweet candidates in the training and test sets.

Tweet ordering. For the second task, we propose a related but more complex task of thread message ordering, based on the sentence ordering task of Barzilay and Lapata (2008) to assess text relatedness. We construct the data by shuffling Twitter threads (containing 3–5 tweets), and assessing the predicted ordering in terms of rank correlation with the original. After removing all duplicates messages, we obtain 7,608 instances for this task.

5 Evaluation Methodology

We provide details of the evaluation methodology in this section.

Morpho-syntax/Sequence Labelling. For POS tagging, we evaluate by 5-fold cross validation using the partitions provided by Kurniawan and Aji (2018). Unlike Kurniawan and Aji (2018) who use macro-averaged F1, we use the standard POS tag accuracy for evaluation. For NER, both corpora (NER UI and NER UGM) are from the news domain. We convert them into `IOB2` format, and reserve 10% of the original training set as a validation set. We evaluate using entity-level F1 over the provided test set.[24] In addition, we conducted our own in-house evaluation of the annotation quality of both datasets by randomly picking 100 sentences and counting the number of annotation errors. We found that NER UI has better quality than NER UGM with 1% vs. 30% errors, respectively. Annotation errors in NER UGM are largely due to low recall, i.e. annotating named entities with the tag `O`.

For dependency parsing we do not apply 5-fold cross-validation for UD-Indo-GSD, as it was released with a pre-defined test set, which allows us to directly benchmark against previous work. UD-Indo-PUD, on the other hand, only includes 1,000 sentences with no fixed test set, so we evaluate via 5-fold cross-validation with fixed splits.[25] Note that the text in UD-Indo-PUD was manually translated from documents in other languages, while UD-Indo-GSD was sourced from texts authored in Indonesian. Additionally, the translation quality of UD-Indo-PUD is low in parts, which impacts on evaluation, as we return to discuss in Section 7. We evaluate both dependency parsing datasets based on the unlabelled attachment score (UAS) and labelled attachment score (LAS).

Semantics. Because the sentiment analysis data is low-resource and imbalanced, we use stratified 5-fold cross-validation, and evaluate based on F1 score. For summarization, on the other hand, we use the canonical splits provided by Kurniawan and Louvan (2018), and evaluate the resulting summary with ROUGE (F1) (Lin, 2004) in the form of three different metrics: R1, R2, and RL.

Discourse Coherence. We do not perform 5-fold cross-validation over NTP for two reasons. First, we need to ensure the distractors in the test set do not overlap with the training or development sets, to avoid possible bias because of dataset artefacts. Second, the size of the dataset in terms of pair-wise labelling is actually four times the reported size (Table 1) as there are three distractors for each thread. We evaluate the NTP task based on accuracy, meaning the random baseline is 25%.

For tweet ordering, we evaluate using Spearman's rank correlation (ρ). Specifically, we average the rank correlation between the gold and predicted order of each thread in the dataset.

6 Comparative Evaluation

6.1 Baselines

Most of our experiments use a BiLSTM with 300d `fastText` pre-trained Indonesian embeddings (Bojanowski et al., 2016) as a baseline. Details of the baselines are provided in Table 2.

For extractive summarization baselines, we use the models of Kurniawan and Louvan (2018) and Cheng and Lapata (2016) as baselines. Kurniawan and Louvan (2018) propose a sentence tagging approach based on a hidden Markov model, while Cheng and Lapata (2016) use a hierarchical LSTM

---

[24]We used the `seqeval` library to evaluate the POS and NER tasks.

[25]The split for cross validation is 70/10/20 for train/development/test, respectively. We first create 5 folds with non-overlapping test partitions, and for each fold we set the first portion of the remaining data as the development (and the rest as training data).

Task	Baselines	BERT models
POS Tagging and NER	**Lample et al. (2016)**[26] A hierarchical BiLSTM + CRF with input: character-level embedding (updated), and word-level `fastText`embedding (fixed), lr: 0.001, epoch:100 with early stopping (patience = 5)	**Fine-tuning:** adding a classification layer for each token, lr: 5e-5, epoch:100 with early stopping (patience = 5)
Dependency parsing	**1. Dozat and Manning (2017)**, Bi-Affine parser, Embedding: `fastText`(fixed) **2. Rahman and Purwarianti (2020)**† **3. Kondratyuk and Straka (2019)**† **4. Alfina et al. (2019)**†	**Dozat and Manning (2017)**, Bi-Affine parser, Embedding: BERT output (fixed)
Sentiment Analysis	**1. 200-d BiLSTM** Embedding: `fastText`(fixed), lr: 0.001, epoch:100 with early stopping (patience = 5) **2. Naive Bayes and Logistic Regression** input: Byte-pair encoding (unigram+bigram)[27]	**Fine-tuning:** Input: 200 tokens; epoch: 20; lr: 5e-5; batch size: 30; warm-up: 10% of the total steps; early stopping (patience = 5); Output layer uses the encoded [CLS]
Summarization	**1. Kurniawan and Louvan (2018)**† **2. Cheng and Lapata (2016)**†	**Liu and Lapata (2019), extractive model,** 20,000 steps, lr: 2e-3, and tokens: 512.[28]
NTP	**200-d BiLSTM (binary-class.)** Embedding: fastText (fixed), lr: 0.001, epoch:100 with early stopping (patience = 20)	**Fine-tuning:** Input: 60 tokens (for 1 single tweet); epoch: 20; learning rate; 5e-5; batch size: 20; warm-up: 10% of the total steps; early stopping (patience = 5); Output layer uses the encoded [CLS]
Tweet Ordering	**Hierarchical 200-d BiLSTMs (multi-class.)** Embedding: fastText (fixed), lr: 0.001, epoch:100 with early stopping (patience = 20)	**Fine-tuning:** Input: 50 tokens (for 1 single tweet); epoch: 20; learning rate; 5e-5; batch size: 20; warm-up: 10% of the total steps; early stopping (patience = 5); BERT fine-tuning is based on the Liu and Lapata (2019) trick (alternated seq.)

Table 2: Comparison of baselines and BERT-based models for all INDOLEM tasks. All listed models were implemented and run by the authors, except for those marked with "†" where the results are sourced from the original paper.

encoder with attention. In addition, we present ORACLE results, obtained by greedily maximizing the ROUGE score between the reference summary and different combinations of sentences from the document. ORACLE denotes the upper bound for the extractive summarization.

For next tweet prediction, we concatenate all premise tweets into a single document, and use a BiL-STM and `fastText`word embeddings to obtain the baseline document encoding. We structure this task as a binary classification where we match the premise with each candidate next tweet. We pick the tweet with the highest probability as the prediction. We use the same BiLSTM to encode the next tweet, and feed the concatenated representations from the last hidden states into the output layer.

For tweet ordering, we use a hierarchical BiLSTM model. The first BiLSTM is used to encode a single tweet by averaging all hidden states. We use the second BiLSTM to learn the inter-tweet ordering. We design the tweet ordering task as a sequence labelling task, where we aim to obtain $P(r|t)$, the probability distribution across rank positions r for a given tweet t. Note that in this experiment, each instance is comprised of 3–5 tweets, and we model the task via multi-classification (with 5 classes/ranks). We perform inference based on $P(r|t)$, where we decide the final rank based on the highest sum of probabilities from the exhaustive enumeration of document ranks.

6.2 BERT Benchmarks

To benchmark INDOBERT, we compare against two pre-existing BERT models: multilingual BERT ("MBERT"), and a monolingual BERT for Malay ("MALAYBERT").[29] MBERT is trained by concatenating Wikipedia documents for 104 languages including Indonesian, and has been shown to be effective for zero-shot multilingual tasks (Wu and Dredze, 2019; Wang et al., 2019c). MALAYBERT is a a publicly available model that was trained on Malay documents from Wikipedia, local news sources, social media, and some translations from English. We expect MALAYBERT to provide better representations than MBERT for the Indonesian language, because Malay and Indonesian are mutually intelligible, with many lexical similarities, but noticeable differences in grammar, pronunciation and vocabulary.

For the sequence labelling tasks (POS tagging and NER), sentiment analysis, NTP, and tweet ordering task, the fine-tuning procedure is detailed in Table 2.

For dependency parsing, we follow Nguyen and Nguyen (2020) in incorporating BERT into the Bi-Affine dependency parser (Dozat and Manning, 2017) by replacing the word embeddings with the corresponding contextualized representations. Specifically, we generate the BERT embedding of the first WordPiece token as the word embedding, and train the BiAffine parser in its default configuration. In addition, we also benchmark against a pre-existing fine-tuned version of MBERT trained over 75 concatenated UD datasets in different languages (Kondratyuk and Straka, 2019).

For summarization, we follow Liu and Lapata (2019) in encoding the document by inserting the tokens [CLS] and [SEP] between sentences. We also apply alternating segment embeddings based on whether the position of a sentence is odd or even. On top of the pre-trained model, we use a second transformer encoder to learn inter-sentential relationships. The input is the encoded [CLS] representation, and the output is the extractive label $y \in \{0, 1\}$ (1 = include in summary; 0 = don't include).

7 Results

Table 3 shows the results for POS tagging and NER. MBERT, MALAYBERT, and INDOBERT perform very similarly over the POS tagging task, well above the BiLSTM baseline. This indicates that all three contextual embedding models are able to generalize well over low-level morpho-syntactic tasks. Given that Indonesian and Malay share a large number of words, it is not surprising that MALAYBERT performs on par with INDOBERT for POS tagging. On the NER tasks, both MALAYBERT and INDOBERT outperform MBERT, which performs similarly to or slightly above the BiLSTM. This is despite MBERT having been trained on a much larger corpus, and having seen many more entities during training. IN-DOBERT slightly outperforms MALAYBERT.

In Table 4, we show that augmenting the BiAffine parser with the pre-trained models yields a strong result for dependency parsing, well above previously-published results over the respective datasets. Over UD-Indo-GSD, INDOBERT outperforms all methods on both metrics. The universal fine-tuning approach (Kondratyuk and Straka, 2019) yields similar performance as BiAffine + `fastText`, while augmenting BiAffine with MBERT and MALAYBERT yields lower UAS and LAS scores than INDOBERT. Over UD-Indo-PUD, we see that augmenting BiAffine with MBERT outperforms all methods including INDOBERT. Note that Kondratyuk and Straka (2019) is trained on the original version of UD-Indo-PUD, and Alfina et al. (2019) is based on 10-fold cross-validation, meaning the results are not 100% comparable.

To better understand why MBERT performs so well over UD-Indo-PUD, we randomly selected 100 instances for manual analysis. We found that 44 out of the 100 sentences contained direct borrowings of foreign words (29 names, 10 locations, and 15 organisations), some of which we would expect to be localized into Indonesian, such as: *St. Rastislav*, *Star Reach*, *Royal National Park Australia*, and *Zettel's Traum*. We also thoroughly examined the translation quality and found that roughly 20% of the sentences

---

[26]The baseline code is available as `chars-lstm-lstm-crf` at `https://github.com/guillaumegenthial/tf_ner`.

[27]We also experimented with simple term frequency, but observed lower performance so omit the results from the paper.

[28]We checkpoint every 2,500 steps, and perform inference over the test set based on the top-3 best checkpoints according to the development set

[29]`https://huggingface.co/huseinzol05/bert-base-bahasa-cased`

Method	POS tagging Acc	NER UGM F1	NER UI F1
BiLSTM-CRF (Lample et al., 2016)	95.4	70.9	82.2
MBERT	**96.8**	71.6	82.2
MALAYBERT	**96.8**	73.2	87.4
INDOBERT	**96.8**	**74.9**	**90.1**

Table 3: Results on POS and NER tasks using accuracy averaged over five folds for POS tagging task, and entity-level F1 over the test set for the NER tasks.

Method	UD-Indo-GSD UAS	LAS	Method	UD-Indo-PUD UAS	LAS
Rahman and Purwarianti (2020)*	82.56	76.04	Alfina et al. (2019)*	83.33	79.39
Kondratyuk and Straka (2019)	86.45	80.10	Kondratyuk and Straka (2019)*	77.47	56.90
BiAffine w/ `fastText`	85.25	80.35	BiAffine w/ `fastText`	84.04	79.01
BiAffine w/ MBERT	86.85	81.78	BiAffine w/ MBERT	**90.58**	**85.44**
BiAffine w/ MALAYBERT	86.99	81.87	BiAffine w/ MALAYBERT	88.91	83.56
BiAffine w/ INDOBERT	**87.12**	**82.32**	BiAffine w/ INDOBERT	89.23	83.95

Table 4: Results for dependency parsing. Methods marked with '*' (from previous work) do not use the same test partition.

Method	Sentiment Analysis (F1)	Method	Summarization (F1) R1	R2	RL
Naive Bayes	70.95	ORACLE	79.27	72.52	78.82
Logistic Regression	72.14	Kurniawan and Louvan (2018)	17.62	4.70	15.89
BiLSTM w/ `fastText`	71.62	Cheng and Lapata (2016)	67.96	61.65	67.24
MBERT	76.58	MBERT	68.40	61.66	67.67
MALAYBERT	82.02	MALAYBERT	68.44	61.38	67.71
INDOBERT	**84.13**	INDOBERT	**69.93**	**62.86**	**69.21**

Table 5: Results over the semantic tasks.

are low-quality translations. For instance, *Ketidaksesuaian data ekonomi dan retorika politik tidak asing, atau seharusnya tidak asing* is not a natural sentence in Indonesian.

For the semantic tasks, INDOBERT outperforms all other methods for both sentiment analysis and extractive summarization (Table 5). For sentiment analysis, the improvement over the baselines is impressive: +13.2 points over naive Bayes, and +7.5 points over MBERT. As expected, MALAYBERT also performs well for sentiment analysis, but substantially lower than INDOBERT. For summarization, MBERT and MALAYBERT achieve similar performance, and only outperform Cheng and Lapata (2016) by around 0.5 ROUGE points. INDOBERT, on the other hand, is 1–2 ROUGE points better.

Lastly, in Table 6, we observe that INDOBERT is once again substantially better than the other models at discourse coherence modelling, despite its training not including next sentence prediction (as per the English BERT). To assess the difficulty of the NTP task, we randomly selected 100 test instances, and the first author (a native speaker of Indonesian) manually predicted the next tweet. The human performance was 90%, lower than the pre-trained language models. For the tweet ordering task, we also assess human performance by randomly selecting 100 test instances, and found the rank correlation score of $\rho = 0.61$ to be slightly higher than INDOBERT. The gap between INDOBERT and the other BERT models was bigger on this task.

Overall, with the possible exception of POS tagging and NTP, there is substantial room for improve-

Method	Next Tweet Prediction (Acc)	Tweet Ordering (ρ)
Random	25.0	0.00
Human (100 samples)	90.0	0.61
BiLSTM w/ `fastText`	73.6	0.45
MBERT	92.4	0.53
MALAYBERT	93.1	0.51
INDOBERT	**93.7**	**0.59**

Table 6: Results for discourse coherence. "Human" is the oracle performance by a human annotator.

ment across all tasks, and our hope is that INDOLEM can serve as a benchmark dataset to track progress in Indonesian NLP.

8 Conclusion

In this paper, we introduced INDOLEM, a comprehensive dataset encompassing seven tasks, spanning morpho-syntax, semantics, and discourse coherence. We also detailed INDOBERT, a new BERT-style monolingual pre-trained language model for Indonesian. We used INDOLEM to benchmark INDOBERT (including comparative evaluation against a broad range of baselines and competitor BERT models), and showed it to achieve state-of-the-art performance over the dataset.

Acknowledgements

References

Mirna Adriani, Ruli Manurung, and Femphy Piscedlo. 2009. Statistical based part of speech tagger for Bahasa Indonesia. In *Proceedings of the 3rd International MALINDO Workshop*.

Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Ika Alfina, Arawinda Dinakaramani, Mohamad Ivan Fanany, and Heru Suhartanto. 2019. Gold standard dependency treebank for Indonesia. In *Proceeding of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), Hakodate, Japan*.

Aristoteles Aristoteles, Yeni Herdiyeni, Ahmad Ridha, and Julio Adisantoso. 2012. Text feature weighting for summarization of document Bahasa Indonesia using genetic algorithm. *IJCSI International Journal of Computer Science Issues*, 9(1):1–6.

Sariwening Azhari and Sarah Lintang. 2020. Indobert: Transformer-based model for Indonesian language understanding. Undergraduate thesis, Universitas Gadjah Mada.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Indra Budi, Stphane Bressan, Gatot Wahyudi, Zainal A. Hasibuan, and Bobby A. A. Nazief. 2005. Named entity recognition for the Indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach. *Discovery Science*, pages 57–69.

Yahya Eru Cakra and Bayu Distiawan Trisedya. 2015. Stock price prediction using linear regression based on sentiment analysis. In *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 147–154.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 484–494.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NeurIPS 2019: Thirty-third Conference on Neural Information Processing Systems*, pages 7057–7067.

Marie-Catherine de Marneffe and Christopher D. Manning. 2016. Stanford typed dependencies manual. Technical report, Stanford University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. 2014. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In *2014 International Conference on Asian Language Processing (IALP)*, pages 66–69.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the 2016 International Conference on Learning Representations*, pages 1–8.

Muhammad Fachri. 2014. Named entity recognition for Indonesian text using hidden Markov model. Undergraduate Thesis.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370.

Nathan Green, Septina Dian Larasati, and Zdenek Zabokrtsky. 2012. Indonesian dependency treebank: Annotation and parsing. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 137–145.

Yohanes Gultom and Wahyu Catur Wibowo. 2017. Automatic open domain information extraction from Indonesian text. In *2017 International Workshop on Big Data and Information Security (IWBIS)*, pages 23–30.

D Gunawan, A Pasaribu, R F Rahmat, and R Budiarto. 2017. Automatic text summarization for Indonesian language using TextTeaser. *IOP Conference Series: Materials Science and Engineering*, 190(1):12048.

Karl Moritz Hermann, Tom Koisk, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 1693–1701.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Mochamad Ibrahim, Omar Abdillah, Alfan F. Wicaksono, and Mirna Adriani. 2015. Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1348–1353.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language. In *ACL 2019 : The 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Mia Kamayani and Ayu Purwarianti. 2011. Dependency parsing for Indonesian. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–5.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pretraining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2779–2795.

Fajri Koto and Gemala Y. Rahmaningtyas. 2017. Inset lexicon: Evaluation of a word list for indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*, pages 391–394.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. to appear. Liputan6: A large-scale Indonesian dataset for text summarization. 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 (AACL-IJCNLP 2020)*.

Fajri Koto. 2016. A publicly available Indonesian corpora for automatic abstractive and extractive chat summarization. In *Proceedings of LREC 2016*.

Kemal Kurniawan and Alham Fikri Aji. 2018. Toward a standardized and more accurate Indonesian part-of-speech tagging. In *2018 International Conference on Asian Language Processing (IALP)*, pages 303–307.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June.

Septina Dian Larasati. 2012. IDENTIC corpus: Morphologically enriched Indonesian-English parallel corpus. In *LREC*, pages 902–906.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *ACL 2020: 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3738.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany.

Andry Luthfi, Bayu Distiawan, and Ruli Manurung. 2014. Building an Indonesian named entity recognizer using Wikipedia and DBPedia. In *2014 International Conference on Asian Language Processing (IALP)*, pages 19–22.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tckstrm, Claudia Bedini, Nria Bertomeu Castell, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Marek Medved and Vít Suchomel. 2017. Indonesian web corpus (idWac). In *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (FAL), Faculty of Mathematics and Physics, Charles University*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. *arXiv preprint arXiv:2003.00744*.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy, July.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Glsen Eryigit, Sandra Kbler, Svetoslav Marinov, and Erwin Marsi. 2005. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Yanuar Nurdiansyah, Saiful Bukhori, and Rahmad Hidayat. 2018. Sentiment analysis system for movie review in Bahasa Indonesia using nave bayes classifier method. *Journal of Physics: Conference Series*, 1008:12011.

Femphy Pisceldo, Ruli Manurung, and Mirna Adriani. 2009. Probabilistic part of speech tagging for Bahasa Indonesia. In *Third International MALINDO Workshop*, pages 1–6.

Valdi Rachman, Septiviana Savitri, Fithriannisa Augustianti, and Rahmad Mahendra. 2017. Named entity recognition on Indonesian twitter posts using long short-term memory networks. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *CoRR, abs/1704.01444, 2017*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Arief Rahman and Ayu Purwarianti. 2020. Dense word representation utilization in Indonesian dependency parsing. *Jurnal Linguistik Komputasional*, 3(1):12–19.

Arief Rahman, Kuncoro Adhiguna, and Ayu Purwarianti. 2017. Ensemble technique utilization for Indonesian dependency parser. *PACLIC*, pages 64–71.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1073–1083.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *NAACL-HLT (1)*, pages 380–385.

F. Tala, J. Kamps, K.E. Mller, and M. de Rijke. 2003. The impact of stemming on information retrieval in Bahasa Indonesia. In *Proceedings of The 14th Meeting of Computational Linguistics in the Netherlands*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5998–6008.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR 2019: 7th International Conference on Learning Representations*.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019c. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5720–5726.

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. *arXiv preprint arXiv:2009.05387*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL HLT 2018: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1112–1122.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 833–844.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS 2019: Thirty-third Conference on Neural Information Processing Systems*, pages 5754–5764.

Daniel Zeman, Jan Haji, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Conference on Computational Natural Language Learning (CoNLL)*, pages 1–21.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.