

# Towards a Standardized Dataset on Indonesian Named Entity Recognition

Siti Oryza Khairunnisa

Aizhan Imankulova

Mamoru Komachi

Tokyo Metropolitan University

6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan

{siti-oryza-khairunnisa, imankulova-aizhan}@ed.tmu.ac.jp  
komachi@tmu.ac.jp

## Abstract

In recent years, named entity recognition (NER) tasks in the Indonesian language have undergone extensive development. There are only a few corpora for Indonesian NER; hence, recent Indonesian NER studies have used diverse datasets. Although an open dataset is available, it includes only approximately 2,000 sentences and contains inconsistent annotations, thereby preventing accurate training of NER models without reliance on pre-trained models.

Therefore, we re-annotated the dataset and compared the two annotations’ performance using the Bidirectional Long Short-Term Memory and Conditional Random Field (BiLSTM-CRF) approach. Fixing the annotation yielded a more consistent result for the organization tag and improved the prediction score by a large margin. Moreover, to take full advantage of pre-trained models, we compared different feature embeddings to determine their impact on the NER task for the Indonesian language.

## 1 Introduction

Named entity recognition (NER) is an essential subtask in natural language processing (NLP). However, NER still suffers from data sparseness for the majority of languages, including Indonesian.

Various Indonesian NER approaches have been proposed, ranging from rule-based methods (Budi et al., 2005) to machine learning-based techniques (Leonandya et al., 2015; Aryoyudanta et al., 2016). The DBpedia and Wikipedia datasets are mainly used for supervised approaches (Alfina et al., 2016; Leonandya et al., 2015; Aryoyudanta et al., 2016; Gunawan et al., 2018). Other datasets include Twitter (Taufik et al., 2016; Wintaka et al., 2019) and conversational datasets from chatbots (Kurniawan and Louvan, 2018), but the sizes of these datasets

are limited. Unfortunately, almost all of the previous studies on Indonesian NER did not release their datasets, which provide essential information for machine learning-based NLP.

One Indonesian NER dataset with human annotation that is openly available is a news dataset from Syaifudin and Nurwidyanoro (2016) (hereinafter referred to as S&N (2016)). However, this dataset exhibits inconsistency problems. In this study, we re-annotated this dataset, thereby developing a more standardized Indonesian NER resource to improve the NLP foundation for the Indonesian language. The most problematic entity is organization, followed by location and person. Certain tokens had been tagged as entities that they were not; for example, the term “DPP” (which means “party’s representative council”) is not an organization name but had been tagged as such.

The most recent Indonesian NER work that used BiLSTM-CRF was conducted by Wintaka et al. (2019) using FastText as the word representation. It has been claimed that FastText offers advantages in handling misspelled words and out-of-vocabulary (OOV) problems (Bojanowski et al., 2017). Therefore, we used BiLSTM-CRF with FastText as our baseline model.

We also experimented with Bidirectional Encoder Representations from Transformers (BERT), a transformer-based language model known to work best in various tasks in NLP as well as NER by acquiring contextual word meanings based on their usage in a sentence (Devlin et al., 2019). For the experiment, we compared three models: the multilingual transformer-based models; mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) and a monolingual BERT for the Indonesian language (IndoBERT) (Wilie et al., 2020). Regarding the limited vocabulary in our low-resourced data, we hypothesized that these embeddings could solve the OOV problem because of the vocabu-

lary coverage in the large-scale data used for pre-training those embeddings.

Our contributions can be summarized as follows:

1. We re-annotated the human-annotated Indonesian NER dataset to improve its consistency and made the dataset publicly available.<sup>1</sup>
2. We analyzed the impact of the data consistency by comparing the performance of NER models trained on the previous and re-annotated datasets. Our dataset significantly improved NER performance.
3. We compared the static and dynamic word embeddings for the Indonesian NER task and showed the impact of different embeddings on the NER model performance.

## 2 Related Works

Several recent NER methods employed the bidirectional neural network and a conditional random field (CRF) as the encoder-decoder layer (Lample et al., 2016; Peters et al., 2018; Akbik et al., 2018). Using contextual information as a representation input to the encoder model improved the score substantially as it helped the model learn the context of the entities. Akbik et al. (2018) introduced Flair embedding, a contextual string embedding approach, and showed that stacking word and character embeddings increased the model’s ability to understand contextual and word-level semantic representations. The use of a Transformer (such as in BERT) also demonstrated significant results for numerous NLP downstream tasks, for example, by BERT, which was proven to work on certain tasks, as well as for NER (Devlin et al., 2019; Conneau et al., 2020). The current state-of-the-art NER model to date was created by fine-tuning a cloze-driven pre-trained bidirectional transformer model (Baevski et al., 2019).

Nevertheless, we focus on low-resourced language NER. The majority of previous studies on low-resourced NER also implemented BiLSTM-CRF as the sequence labeling method and experimented with input representation (Pham and Le-Hong, 2018; Poostchi et al., 2018; Singh et al., 2019). BERT has also been employed in several low-resource languages, including Bulgarian (Marinova, 2019), Arabic (Antoun et al., 2020), and Basque (Agerri et al., 2020).

<sup>1</sup><https://github.com/khairunnisaor/idner-news-2k>

Deep learning has recently been used in Indonesian NER research. The most widely used method is the BiLSTM algorithm (Huang et al., 2015; Lample et al., 2016). Various input representation methods have been applied, such as convolutional neural networks (CNNs) for word  $n$ -gram representation (Gunawan et al., 2018) and pre-trained word embeddings with part-of-speech (PoS) tags (Hoesen and Purwarianti, 2018). In exploring the OOV problem in conversational text, Kurniawan and Louvan (2018) also employed BiLSTM-CRF without including any pre-trained word representation. In recent work by Wintaka et al. (2019), the same neural sequence labeling model was implemented, and pre-trained FastText Indonesian word embedding was applied as the input. In work similar to ours, Leonandya and Ikhwantri (2019) investigated the impact of language model pre-training on the NER task. However, their conversational texts data is not publicly available, and therefore their study is not replicable. The latest Indonesian NER work is from Wilie et al. (2020), who built and fine-tuned an Indonesian BERT (IndoBERT) pre-trained model for twelve NLP tasks, including NER.

Previous work using the same dataset was conducted by S&N (2016) for a quotation identification task. The dataset was constructed from three Indonesian online news sites, namely Kompas<sup>2</sup>, Tempo<sup>3</sup>, and TribunNews<sup>4</sup>. The topics covered by this dataset mainly concern politics, society, and economics. In this task, the data were labeled for quotation identification. However, the NER data were manually tagged as well, as they were used in preprocessing for the quotation identification task. In this study, we focused on the NER task and re-annotated the data because of the inconsistency.

## 3 Methodology

### 3.1 Inconsistency of Existing Dataset

We used an open dataset released by S&N (2016), which is available on GitHub<sup>5</sup>. However, we found that several tokens in the dataset were not tagged correctly. For example, tokens of certain organizations and persons were not tagged or were tagged incorrectly. Table 1 shows an examples of inconsistency in the annotation. The three sentences

<sup>2</sup><https://www.kompas.com/>

<sup>3</sup><https://www.tempo.co/>

<sup>4</sup><https://www.tribunnews.com/>

<sup>5</sup><https://github.com/yusufsyafudin/Indonesia-ner>

Sentence 1												
President Joko Widodo met the chairman of Gerindra Prabowo Subianto at Istana Bogor												
Indonesian	Presiden	Joko	Widodo	bertemu	Ketua	Umum	Gerindra	Prabowo	Subianto	di	Istana	Bogor
English translation	President	Joko	Widodo	met	chairman	general	Gerindra	Prabowo	Subianto	at	palace	Bogor
S&N (2016)	O	B-PER	I-PER	O	B-PER	I-PER	I-PER	I-PER	I-PER	O	B-LOC	I-LOC
Ours	O	B-PER	I-PER	O	O	O	B-ORG	B-PER	I-PER	O	B-LOC	I-LOC

  

Sentence 2												
Politician PDI Perjuangan Guruh Soekarnoputra visited the chairman of Gerindra Party Suhardi												
Indonesian	Politikus	PDI	Perjuangan	Guruh	Soekarnoputra	menjenguk	Ketua	Umum	Partai	Gerindra	Suhardi	
English translation	Politician	PDI	Perjuangan	Guruh	Soekarnoputra	visited	chairman	general	Party	Gerindra	Suhardi	
S&N (2016)	O	B-ORG	I-ORG	B-PER	I-PER	O	O	O	O	O	B-PER	
Ours	O	B-ORG	I-ORG	B-PER	I-PER	O	O	O	B-ORG	I-ORG	B-PER	

  

Sentence 3												
Vice chairman of Gerindra Party Edy Prabowo stated that his party would not be hurt												
Indonesian	Wakil	Ketua	Umum	Partai	Gerindra	Edy	Prabowo	menyatakan	partainya	tak	akan	sakit hati
English translation	Vice chairman	chairman	general	Party	Gerindra	Edy	Prabowo	stated	his party	not	would	hurt
S&N (2016)	O	O	O	B-ORG	I-ORG	B-PER	I-PER	O	O	O	O	O
Ours	O	O	O	B-ORG	I-ORG	B-PER	I-PER	O	O	O	O	O

Table 1: Examples of tags before (S&N (2016)) and after (ours) re-tagging. The red tokens indicate the difference after re-tagging. The blue tokens represent consistent annotation between S&N (2016) and ours. Tag prefix meanings: B indicates the entity’s first word, whereas I indicates the second and remaining part of the entity.

Data Split	Sentences	Tokens
Train	1,464	30,248
Development	367	7,863
Test	509	10,588
Total	2,340	48,699

Table 2: Data statistics.

		Our Tags				# of tags
		LOC	ORG	PER	O	
S&N (2016) Tags	LOC	1,153	26	4	344	1,527
	ORG	5	1,562	2	78	1,647
	PER	4	3	2,317	39	2,363
	O	91	701	127	42,241	43,160
# of tags		1,253	2,292	2,450	42,702	

Table 3: Confusion matrix of our re-annotation from S&N (2016). The number of tags is represented at the token level. The first column indicates the entity’s previous tag and the header denotes our tag in the re-annotation. LOC: location; ORG: organization; PER: person; O: other.

have a pattern of [title] [organization] [person]. In Sentence 1, S&N (2016) tagged all tokens “Ketua Umum” (title), “Gerindra” (organization name) and “Prabowo Subianto” (person name) as person names. In the red part of Sentence 2, S&N (2016) only tagged “Suhardi” (person name) as a person name and did not tag “Partai Gerindra” as an organization name. In the blue part of Sentences 2 and 3, S&N (2016) did not tag the titles (“Politikus” and “Wakil Ketua Umum”) as entities but did tag the organization and person name correctly. These sentences demonstrate that there were three different annotations for the same sentence pattern. As this phenomenon appeared several times, we

examined the dataset and checked the annotations one by one.

### 3.2 Dataset Re-annotation

To address these problems, we asked three native speakers to re-annotate the data manually. Although S&N (2016) covers five entities, we only re-annotated only three entities that are commonly used in the NER task, namely location, organization, and person. We omitted the other two, time and quantity, because we wanted to build a model with a strong foundation for recognizing ambiguous nouns. As most time and quantity entities are written in numeric form, they are easy to be recognized by a well-developed NER model. In the experiment, we considered just the three entities in both datasets so that the results would be reasonably comparable.

We calculated the inter-annotator agreement of the three annotators and obtained an agreement score of 0.92 using Fleiss’ kappa (Fleiss, 1971), which indicates a high agreement and good reliability (Artstein and Poesio, 2008). In this study, we used the same split as that used in S&N (2016). However, owing to the absence of the development set as in theirs, we randomly sampled data from the training set to constitute the development set, as indicated in Table 2.

### 3.3 Our Annotation Guidelines

To clearly differentiate how we annotated each entity, here we provide the guidelines we used in re-annotating the dataset.

- **Location**: indicates the name of a location name where activities or events happened semantically. Such an entity is usually preceded by a location preposition, namely “di” (at),

“*ke*” (to), or “*dari*” (from). Specific location names such as a country or city name (e.g., Indonesia in “Indonesia is one of the largest countries”) when not used contextually as a location would not be annotated as a location. An organization name (e.g., university or office), conversely, is sometimes used as a location name when the sentence refers to its building or location. In this case, we annotate the entity as a location name.

- **Organization:** indicates an organization’s name. The name of the organization is usually an official institution that is legally registered.
- **Person:** identifies a person’s name. Any form of the person’s name—full, nickname, or abbreviation—is annotated as one name. For example, “*Abu Rizal Bakrie*” is the full name of a person, who may also be mentioned as “*Ical*” (nickname) or “*ABR*” (abbreviation). A person’s title, such as “*Pak*” (Mr.) in “*Pak Ical*” (Mr. Ical) is not included in the person’s name; it is annotated as “*Pak [Ical]<sub>B-PER</sub>*”, not as “*[Pak]<sub>B-PER</sub> [Ical]<sub>I-PER</sub>*”.
- An organization or person name that is sometimes written in full may, at other times, be written in its abbreviated form. When both forms appear, the annotation will be separated into two entities. For example, the sentence, “*Universitas Gadjah Mada (UGM) berlokasi di Yogyakarta.*” (Gadjah Mada University (UGM) is located in Yogyakarta) is annotated, as shown below:

[Universitas]<sub>B-ORG</sub> [Gadjah]<sub>I-ORG</sub> [Mada]<sub>I-ORG</sub>  
 ([UGM]<sub>B-ORG</sub>) berlokasi di [Yogyakarta]<sub>B-LOC</sub>

For ambiguous entities, the tags were determined according to the word’s contextual use; that is, whether it is used as a location name or organization name (these two entities are generally the most confusing). When there was a disagreement, the tag chosen by the majority of the taggers was determined as the final tag.

Table 3 presents the confusion matrix of our re-annotation of S&N (2016). The number of tags for the organization and person entities increased after the re-tagging. Meanwhile, the number of location entities was reduced from approximately 1,500 to 1,200, and the other (O) entity was reduced by almost 500 tags. This indicates that 20% of the

location tags were incorrect, and almost 500 tokens were not tagged. Comparing our annotation, we calculated the percent difference by dividing the difference in the number of tags by the total number of tokens.

## 4 Experiment

### 4.1 NER Methods

BiLSTM-CRF is a deep learning algorithm introduced by Huang et al. (2015) and has mainly been used for the NER task owing to its ability to solve sequence tagging problems. Following its successes in dealing with English NER tasks (Lample et al., 2016; Akbik et al., 2018; Ma and Hovy, 2016), BiLSTM-CRF was also implemented in recent Indonesian NER; relatively good results were obtained compared to those of rule-based and former machine learning approaches (Hoesen and Purwarianti, 2018; Kurniawan and Louvan, 2018; Wintaka et al., 2019). We employed a method used by Wintaka et al. (2019) as our baseline. They used FastText (Bojanowski et al., 2017), a pre-trained word embedding with sub-word features, as the input representation for the BiLSTM-CRF.

In addition to FastText, we also used some pre-trained multilingual and monolingual models as the input representation for BiLSTM-CRF. For multilingual models, we applied multilingual BERT (mBERT) by Devlin et al. (2019) and XLM-R by Conneau et al. (2020), and for the monolingual model, we applied the Indonesian monolingual BERT pre-trained model IndoBERT by Wilie et al. (2020). To compare the use of BERT with the feature representation approach, we also investigated the potential benefit of using a fine-tuning approach with all BERT pre-trained models.

### 4.2 Settings

We used the implementation of the BiLSTM-CRF approach provided by Flair NLP<sup>6</sup>, a simple framework for sequence labeling tasks (Akbik et al., 2018). We identified three entity types, namely location (LOC), organization (ORG), and person (PER), and used the IOB format defined by Tjong Kim Sang (2000). We conducted five experiments for each model and calculated the average scores. We applied two approaches for the NER task implementation.

First, in the feature-based approach, we used BiLSTM-CRF and experimented with several word

<sup>6</sup><https://github.com/flairNLP/flair>



Dataset	Annotation	Overall Scores			LOC			ORG			PER		
		P	R	F	P	R	F	P	R	F	P	R	F
S&N	S&N	<u>69.20</u>	84.55	<u>76.11</u>	74.82	<u>85.78</u>	79.92	<u>45.54</u>	83.02	<u>58.82</u>	83.60	84.86	84.23
Ours	S&N	66.39	88.28	75.79	<u>79.05</u>	85.57	82.18	39.64	88.95	54.84	84.95	88.60	86.74
S&N	Ours	89.27	80.06	84.41	80.70	85.19	82.88	88.12	71.21	78.77	92.13	85.91	88.91
Ours	Ours	<b>92.23</b>	<b>89.52</b>	<b>90.85</b>	<b>89.02</b>	<b>88.52</b>	<b>88.76</b>	<b>89.13</b>	<b>88.13</b>	<b>88.63</b>	<b>95.50</b>	<b>90.83</b>	<b>93.10</b>

Table 4: Baseline model comparison of S&N (2016)’s and our annotation performance. The bold scores show the best score for both models when tested on our test set, and the underlined scores present the best score when tested on S&N (2016)’s test set.

Models	Features	Overall Scores			F <sub>1</sub> Scores for Each Tag		
		P	R	F	LOC	ORG	PER
BiLSTM-CRF	FastText (baseline)	92.23	89.52	90.85±0.08	88.76	88.63	93.10
	mBERT	90.11	89.80	89.95±0.05	87.30	86.54	93.20
	XML-R	91.12	94.05	92.56±0.54	89.01	88.95	96.25
	IndoBERT	<b>94.93</b>	<b>94.87</b>	<b>94.90±0.31</b>	<b>89.84</b>	<b>92.99</b>	<b>97.49</b>
Fine-tune mBERT	–	89.54	92.33	90.91±0.51	<b>86.73</b>	87.78	94.11
Fine-tune XML-R	–	91.60	<b>94.76</b>	<b>93.15±0.23</b>	85.59	90.46	<b>97.41</b>
Fine-tune IndoBERT	–	<b>91.66</b>	94.64	93.13±0.31	83.84	<b>91.81</b>	96.83

Table 5: BiLSTM-CRF model performance for contextual embedding experiment.

embeddings as the input representation, namely FastText (Grave et al., 2018), mBERT (Devlin et al., 2019)<sup>7</sup>, XML-R (Conneau et al., 2020), and IndoBERT (Wilie et al., 2020); these are shown in Table 5. The parameter settings were as follows: a learning rate of 0.1, a dropout of 0.5, a mini-batch size of 32, a maximum of 200 epochs, one BiLSTM hidden layer, and 256 BiLSTM hidden units. The implementation used an early stopping method, where the model would stop training when the loss score did not improve in a sequence of five epochs.

Second, we fine-tuned all of the BERT models for the NER task. The output layer used for fine-tuning the BERT models was the standard softmax layer. For the fine-tuning settings, we set the batch size to 32, the number of epochs to five, and the learning rate to [3e-5, 5e-5, 7e-5] respectively.

## 5 Results

**Annotation performance.** Table 4 presents a comparison of the annotation performance between S&N (2016) and our annotation. We conducted a cross test for each model on both annotations’ test sets with the aim of observing both models’ performance and comparing them on an equal footing by testing them on the same annotations. For the S&N

(2016), our baseline model provided an F<sub>1</sub> score of 76.11. Our annotation obtained a higher score, 84.41. Particularly in the organization tag, the score jumped relatively high by almost 20 points.

The phenomenon whereby many organization tokens were not tagged accounted for a sharp decrease in the F<sub>1</sub> score of the ORG tag in the S&N (2016) dataset (ORG: 54.84, LOC: 82.18 and PER: 86.74). Using our test data, we obtained a very high overall score with more consistent performance of the organization tag, as indicated by all three tags sharing relatively similar scores.

**Multilingual vs. monolingual pre-trained models.** Table 5 presents the results using pre-trained BERT models for our Indonesian NER task. The best result was obtained from the feature-based approach with the IndoBERT pre-trained model as the input representation for the BiLSTM-CRF architecture. Fine-tuning those models resulted in scores slightly under those of the feature-based approach. Regarding multilingual model performance, the XML-R outperformed the mBERT model owing to its more extensive unsupervised multilingual data when pre-trained (Conneau et al., 2020).

The monolingual model performed better than the multilingual model when used as the feature representation for BiLSTM-CRF; the multilingual models are better when fine-tuned. Wilie et al.

<sup>7</sup>[https://huggingface.co/transformers/pretrained\\_models.html](https://huggingface.co/transformers/pretrained_models.html)

Sentence 1								
Joko Widodo met Gerindra's Chairman Prabowo Subianto								
Indonesian	Joko	Widodo	bertemu	Ketua	Umum	Gerindra	Prabowo	Subianto
English translation	Joko	Widodo	met	chairman	general	Gerindra	Prabowo	Subianto
S&N (2016) annotation	B-PER	I-PER	O	B-PER	I-PER	I-PER	I-PER	I-PER
S&N (2016) FastText	B-PER	I-PER	O	O	O	O	B-PER	I-PER
Our annotation	B-PER	I-PER	O	O	O	B-ORG	B-PER	I-PER
Our FastText	B-PER	I-PER	O	O	O	B-ORG	B-PER	I-PER
Sentence 2								
Required by the Ministry of Law and Human Rights								
Indonesian	disyaratkan	oleh	Kementerian	Hukum	dan	Hak	Asasi	Manusia
English translation	required	by	ministry	law	and	right	basic	human
S&N (2016) annotation	O	O	O	O	O	O	O	O
S&N (2016) FastText	O	O	B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG
Our annotation	O	O	B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG
Our FastText	O	O	B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG

Table 6: Examples of errors in prediction comparing S&N (2016) and our annotation when trained using the baseline BiLSTM-CRF model. Red indicates incorrect tokens and blue indicates the correct ones.

(2020) stated that the XLM-R model might achieve a better result on the NER task as entity names usually come from English or other languages. However, most of our dataset’s entity names are in Indonesian, and hence, the IndoBERT pre-trained model supports this condition better. Additionally, mBERT and XLM-R use WordPiece (Wu et al., 2016) and SentencePiece (Kudo, 2018) tokenization respectively, whereby longer tokens are split into more common tokens. Multilingual models contain many languages in their vocabularies. When the pre-trained model is frozen for feature representation use, it is possible that some Indonesian sub-words are biased because they share representations with other language’s sub-words. The IndoBERT was trained on the Indo4B dataset, which also contains Indonesian news corpora (Wilie et al., 2020). We hypothesize that the BiLSTM-CRF architecture fits better with our sequence classification task, supported by the rich Indonesian vocabularies covered by IndoBERT and the domain similarity between the Indo4B and our dataset. This is reflected by the very high organization scores obtained from IndoBERT models in both approaches.

## 6 Discussion

Table 6 presents some examples of errors by the model trained on S&N (2016) and on our annotation. In Sentence 1, the words “Ketua Umum Gerindra” were tagged as part of a person’s name, although they are not. The S&N (2016) model identified “Prabowo Subianto” correctly as a person’s name but did not tag “Gerindra” as an organization name. Meanwhile, our model correctly tagged “Ke-

tua Umum” as not being any entity and “Gerindra” as an organization. In Sentence 2, the S&N (2016) annotation did not tag any of the words. However, the baseline model trained on S&N (2016) and on our annotation recognized the tokens as an organization name in all cases. These examples demonstrate that errors in an annotation could affect the prediction performance and decrease the model’s score.

## 7 Conclusions and Future Work

We re-annotated the human-annotated Indonesian NER dataset to produce a more consistent annotation in the Indonesian NER task. This re-annotation obtained an  $F_1$  score of 90.85 when using the baseline BiLSTM-CRF model was used with FastText. Our implementation also demonstrated that the use of pre-trained transformer-based language models, both the multilingual and the monolingual models, yielded better prediction results. Although the performance of Indonesian NER using either BiLSTM-CRF or fine-tuning depends on the pre-trained language model, we found that IndoBERT works best when using BiLSTM-CRF architecture, compared to the fine-tuning approach.

In the future, we plan to address word ambiguity in Indonesian by creating a gazetteer to add more supervision and perform distant supervised learning to aid the model in differentiating a word to be classified as each entity as in Nooralahzadeh et al. (2019). Also, we would like to work on other techniques such as transferring knowledge using a teacher-student learning from a high resource language such as English to a low-resource, such as Indonesian (Wu et al., 2020; Sun et al., 2019).

## References

- Rodrigo Agerri, Iñaki S. Vicente, Jon A. Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. [Give your text representation models some love: the case for Basque](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4781–4788.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. [Contextual string embeddings for sequence labeling](#). In *Proceedings of the COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.
- Ika Alfina, Ruli Manurung, and Mohamad I. Fanany. 2016. [DBpedia entities expansion in automatically building dataset for Indonesian NER](#). In *Proceedings of the 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 335–340.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. [AraBERT: Transformer-based model for Arabic language understanding](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.
- Ron Artstein and Massimo Poesio. 2008. [Survey article: Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Bayu Aryoyudanta, Teguh B. Adji, and Indriana Hidayah. 2016. [Semi-supervised learning approach for Indonesian named entity recognition \(NER\) using co-training algorithm](#). In *Proceedings of the 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 7–12.
- Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. [Cloze-driven pretraining of self-attention networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369.
- 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.
- Indra Budi, Stéphane 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](#). In *Discovery Science*, pages 57–69.
- 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.
- 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.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. [Learning word vectors for 157 languages](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487.
- William Gunawan, Derwin Suhartono, Fredy Purnomo, and Andrew Ongko. 2018. [Named-entity recognition for Indonesian language using bidirectional LSTM-CNNs](#). In *Proceedings of the 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018): Empowering Smart Technology in Digital Era for a Better Life*, volume 135, pages 425 – 432.
- Devin Hoesen and Ayu Purwarianti. 2018. [Investigating Bi-LSTM and CRF with POS tag embedding for Indonesian named entity tagger](#). In *Proceedings of the 2018 International Conference on Asian Language Processing (IALP)*, pages 35–38.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. [Bidirectional LSTM-CRF Models for Sequence Tagging](#). *arXiv e-prints*, page arXiv:1508.01991.
- Taku Kudo. 2018. [Subword regularization: Improving neural network translation models with multiple subword candidates](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.
- Kemal Kurniawan and Samuel Louvan. 2018. [Empirical evaluation of character-based model on neural named-entity recognition in Indonesian conversational texts](#). In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 85–92.
- 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.

- Rezka A. Leonandya, Bayu Distiawan, and Nursidik H. Praptono. 2015. [A semi-supervised algorithm for Indonesian named entity recognition](#). In *Proceedings of the 3rd International Symposium on Computational and Business Intelligence (ISCBI)*, pages 45–50.
- Rezka A. Leonandya and Fariz Ikhwantri. 2019. [Pre-trained language model transfer on neural named entity recognition in Indonesian conversational texts](#). In *Proceedings of the 33rd Pacific Asia Conference on Language, Information, and Computation*, pages 104–113.
- Xuezhe Ma and Eduard Hovy. 2016. [End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.
- Iva Marinova. 2019. [Evaluation of stacked embeddings for Bulgarian on the downstream tasks POS and NERC](#). In *Proceedings of the Student Research Workshop Associated with Recent Advances in Natural Language Processing (RANLP) 2019*, pages 48–54.
- Farhad Nooralahzadeh, Jan Tore Lønning, and Lilja Øvrelid. 2019. [Reinforcement-based denoising of distantly supervised NER with partial annotation](#). In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 225–233.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). 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 2227–2237.
- Thai-Hoang Pham and Phuong Le-Hong. 2018. [End-to-end recurrent neural network models for Vietnamese named entity recognition: Word-level vs. character-level](#). In *Proceedings of the 2017 International Conference of the Pacific Association for Computational Linguistics*, pages 219–232.
- Hanieh Poostchi, Ehsan Z. Borzeshi, and Massimo Piccardi. 2018. [BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4427–4431.
- Oyesh M. Singh, Ankur Padia, and Anupam Joshi. 2019. [Named entity recognition for Nepali language](#). In *Proceedings of the 5th International Conference on Collaboration and Internet Computing (CIC)*, pages 184–190.
- Linghao Sun, Huixiong Yi, Yong Liu, Huanhuan Chen, and Chunyan Miao. 2019. [Back Attention Knowledge Transfer for Low-resource Named Entity Recognition](#). *arXiv e-prints*, page arXiv:1906.01183.
- Yusuf Syaifudin and Arif Nurwidyantoro. 2016. [Quotations identification from Indonesian online news using rule-based method](#). In *Proceedings of the 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 187–194.
- Natanael Taufik, Alfian F. Wicaksono, and Mirna Adriani. 2016. [Named entity recognition on Indonesian microblog messages](#). In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*, pages 358–361.
- Erik F. Tjong Kim Sang. 2000. [Text chunking by system combination](#). In *Proceedings of the Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, pages 151–153.
- Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, X. Li, Zhi Yuan Lim, S. Soleman, R. Mahendra, Pascale Fung, Syafri Bahar, and A. 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*.
- Deni C. Wintaka, Moch A. Bijaksana, and Ibnu Asror. 2019. [Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF](#). In *Proceedings of the 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019): Enabling Collaboration to Escalate Impact of Research Results for Society*, volume 157, pages 221 – 228.
- Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. [Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. [Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation](#). *arXiv e-prints*, page arXiv:1609.08144.