Master's Thesis

Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language

Siti Oryza Khairunnisa

September 23, 2021

Graduate School of Systems Design
Tokyo Metropolitan University

A Master's Thesis
submitted to Graduate School of Systems Design,
Tokyo Metropolitan University
in partial fulfillment of the requirements for the degree of
Master of Computer Science

Siti Oryza Khairunnisa

Thesis Committee:

Associate Professor Mamoru Komachi	(Supervisor)
Professor Toru Yamaguchi	(Co-supervisor)
Professor Yasufumi Takama	(Co-supervisor)

Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language*

Siti Oryza Khairunnisa

Abstract

Natural language processing (NLP) in the Indonesian language has been through significant development in recent years. Named entity recognition (NER) is a fundamental task in information extraction, yet it still lacks extensive and standardized corpora publicly available in the Indonesian language. A small dataset is available, but it contains inconsistent annotations. Thereby, we did a re-annotation of the dataset to be more consistent in performing better training for the models. We examine the performance of both annotations by applying Bidirectional Long Short-Term Memory (BiLSTM) and Conditional Random Field (CRF) as our baseline. We obtained a positive result for the organization tag and improved the overall score quite significantly.

In order to address the data sparseness problem, we exploit the monolingual and multilingual pre-trained language models (PLM) such as IndoBERT and XLM-RoBERTa to help the NER model improve the performance by understanding each word's context. Although both increase the model performance vastly, our thorough analysis shows that the IndoBERT is too context-sensitive, while the multilingual PLMs highly depend on the morphological information and the subwords from their multilingual vocabularies. Furthermore, we explore the use of cross-lingual transfer to utilize the availability of NER corpora in high-resource languages. We acquire two cross-lingual transfer methods, namely data and model

*Master's Thesis, Graduate School of Systems Design Science,
Tokyo Metropolitan University, September 23, 2021.

transfer, with English, Spanish, Dutch, and German as the source languages to the target Indonesian language.

Keywords:

named entity recognition, Indonesian language, cross-lingual transfer learning, natural language processing

Contents

List of Figures

1 Introduction

The Indonesian language is one of the low-resource languages in NLP, specifically with the lack of extensive public corpora for the NER task. Many approaches–ranging from rule-based [10] to machine learning-based [6,40] methods–have been employed to build the NER models. Most studies used DBpedia and Wikipedia for creating the datasets of their approach [4, 6, 22, 40]. Other datasets such as conversational texts [34] from chatbots and Twitter [60, 70] are also applied, although their size is limited. However, most previous Indonesian NER studies did not publish their datasets, which has an essential role in developing machine learning-based NLP.

An open dataset with human annotation for Indonesian NER published by Syaifudin and Nurwidyantoro (hereinafter referred to as S&N (2016)) [59] in news domain is available with about 2,000 sentences. However, it undergoes an inconsistency problem. Meanwhile, our first goal is to develop a standardized dataset that is available online by enhancing the existing dataset since training a model in a noisy data would result in a poor model and produce mispredictions. Therefore, we want to increase the number of reliable public Indonesian NER dataset by improving the annotation quality of the existing dataset. We found that the organization entity has the most ambiguity and followed by location and person entities. One example of the inconsistency is some tokens were tagged as an organization where they were not; the term "DPP" (meaning in English: party's representative council") is classified as an organization, although it is not.

In the recent decade, the neural network approach was widely used in the NLP field. Lample et al. [37] introduced Bidirectional LSTM and CRF to handle sequence tagging problem in English NER. For the input representation, they use a character-based representation model to capture the orthographic sensitivity by learning the character-level feature and combining it with pre-trained word

embeddings to learn the word order distributions over sentences to capture the distributional sensitivity. The orthographic and distributional representations help the model to handle the out-of-vocabulary (OOV) problem and misspelled words. FastText is a more recent word representation with similar ability to handle both orthographic and distributional sensitivity [8]. Wintaka et al. [70] applied the FastText as input representation of the BiLSTM-CRF for an Indonesian NER task on a Twitter dataset and benefits the OOV issue in the Indonesian language. In this study, we want to include the ability of dynamic word embeddings to acknowledge the contextual meaning of words in a sentence by exploiting the use of various PLMs, both in monolingual and multilingual ones.

PLM works effectively in various downstream task in NLP [3, 15, 49]. Despite its feature in understanding the contextual knowledge from training corpora, it can learn representations from multiple languages and helps the low-resource languages to learn a shared knowledge from the richer languages [15]. Therefore, we further investigate how monolingual and multilingual PLMs benefit our issues in Indonesian NER since the monolingual PLM has more knowledge in the language specific features. We involved Bidirectional Encoder Representation from Transformers (BERT), a transformer-based language model that obtains the word representation contextually based on the sentence. We compared three models for the experiment, two models for the multilingual transformer-based models–mBERT [15] and XLM-R [13]–and a monolingual BERT for the Indonesian language–IndoBERT [69]. The use of various contextual embeddings could solve the OOV problem in our limited vocabulary in the dataset since these embeddings used large-scale unlabeled corpora during pre-training.

We also experimented with an unsupervised cross-lingual transfer learning to leverage the knowledge from the high to the low-resource languages. We examined both single-source and multi-source cross-lingual transfer learning with English as the single-source language and English, Spanish, Dutch, and German as the multi-source languages to the target language, Indonesian. Both of our cross-lingual transfer approaches show a competitive results for the NER task in the Indonesian language without relying on a high-resource Indonesian labeled dataset.

2 Related Works

We divide the prior studies into monolingual NER for the supervised approach and cross-lingual transfer learning for the approach of knowledge transfer from the high-resource languages.

2.1 Monolingual NER

NER is a fundamental task in NLP that extracts pre-defined entity names from an unlabeled corpora [76]. It can be any entity that we want to identify such as person, organization, date, location, currency, etc. NER is categorized as a sequence labeling task that treats the words in a sentence as a sequence and labels each of them to the pre-defined types [24]. Intuitively, the task classifies the words by recognizing if the current word is an entity or not, and then defining the type of the entity. A standard BIO (beginning - inside - outside) format is used to determine the order of the words in an entity [54, 63]. The first word of an entity is labeled with a prefix "B" that indicates beginning and the rest is prefixed with "I" that indicates inside. The entity type is written after the BIO with its abbreviation form, e.g., B-PER.

One of the recent NER methods is a bidirectional neural network with CRF as its decoder layer [3, 37, 49]. Contextual word representations are recently used as an input for the encoder model. Flair embeddings is a dynamic language model based on a recurrent neural network, where the sequence of the characters in the sentence represents each word [3]. Akbik et al. [3] demonstrated that stacking word and character embeddings enhanced the model to understand each word's context better. The latest language model, such as Transformer, also showed significant improvements for various NLP downstream tasks. For example, BERT performed very well on many downstream tasks, including NER [13, 15]. The

current state-of-the-art NER model to date is LUKE, a language model with entity-aware self-attention and contextualized entity representation [77]. LUKE works not only for NER, but also other entity-related tasks such as entity typing, relation classification, and question answering. As shown in prior studies that stacking word embeddings could result in some gain for the NER tasks [3, 44], a recent work on six structured prediction tasks also showed a method to find the best concatenation of several word embedding types to improve the performance of a NER task [68].

Early Indonesian NER models adopted a rule-based approach with supervision from contextual dictionary, morphological feature, and part-of-speech (POS) to perform the NER task [10]. Budi and Bressan examined an association rule mining approach with a thorough explanation about the characteristics of Indonesian Language for NER and co-reference resolution tasks [9]. In the last decade, machine learning approach is widely explored for developing Indonesian NER models. Several statistical machine learning approaches were investigated, namely support vector machine (SVM) [35], CRF with gazetteer and POS information [46], and used a semi-supervised method to increase the number of training data in the low-resource settings [41].

In terms of deep learning approach for Indonesian NER studies, the BiLSTM has been widely used [26, 37]. Various input representation methods have been applied, such as convolutional neural networks (CNNs) for word n-gram representation [22] and pre-trained word embeddings with POS tags [25]. In exploring the OOV problem in conversational text, Kurniawan and Louvan [34] also employed BiLSTM-CRF without including any pre-trained word representation. In recent work by Wintaka et al. [70], the same neural sequence labeling model was implemented, and pre-trained FastText Indonesian word embedding was applied as the input. In a work similar to ours, Leonandya and Ikhwantri [42] 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 works exploiting PLMs include IndoBERT [69] and IndoLEM [31], both are Indonesian PLMs evaluated over a set of NLP tasks and dataset in Indonesian language, including NER. Besides the high use of PLMs for Indonesian NER, Fu et al. [19] constructed a hierarchi-

cal structured-attention-based model (HSA) to exploit deeper features from the morphological-rich characteristic of the Indonesian language.

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*, Tempo†, and TribunNews‡. 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 our study, we focused on the NER task and re-annotated the data because of the inconsistency, as Koto et al. [31] also stated that this dataset contains about 30% errors in the annotation.

2.2 Cross-lingual Transfer Learning in NER

Cross-lingual transfer learning is an alternative to improve the lack of available datasets in low-resource NLP tasks by exploiting the knowledge available in high-resource languages [14]. Various studies has been conducted in many NLP subtasks such as neural machine translation [29, 30], grammar error correction [78], and POS tagging [17]. Performing cross-lingual transfer can be divided into two ways, which are 1) data transfer [28, 47, 75] and 2) model transfer based method [71–73]. In data transfer based method, the high-resources dataset such as English is translated into low-resource target languages or vice versa, and extended the use of state-of-the-art English model of the designated task.

Xie et al. [75] conducted a data transfer for cross-lingual NER using monolingual pre-trained model of each source and target language and placed them in the same vector space to obtain the word-to-word translation. Later on, Chaudary et al. [12] further investigated the impact of a small extra annotation for the method by performing annotation only for uncertain tags on the predicted pseudodata in the target language to lessen the effort of manual annotators. Instead of examining a word-to-word translation, Sun et al. [58] proposed a back trans-

*https://www.kompas.com/
†https://www.tempo.co/
‡https://www.tribunnews.com/

lation method, where the target language is translated to the source language and accessing the BiLSTM output states from the translated pre-trained English NER model to be applied on the target language NER model. Jain et al. [28] explored the cross-lingual transferability for NER by leveraging machine translation systems and matching entities based on orthographic and phonetic similarity.

The most common method in model transfer in direct transfer from a NER model of the source language to do inference in the target language in a zero-shot scenario [50, 73]. An enhanced meta-learning algorithm is proposed for cross-lingual NER by computing sentence similarity to construct multiple pseudo-NER task [72]. Wu et al. [71] performed a teacher-student learning, where it highly depends on multilingual PLM to create a teacher model that produce a pseudo-data to train the student model for the target language. Regardless of the extensive ways in transferring knowledge from a high-resource language, some studies also showed positive results in conducting cross-lingual transfer from multi-source languages without any parallel corpora [16, 45, 66]. Rahimi et al. [53] examined a few-shot learning for NER and thoroughly investigated the mistakes and language specific transfer-errors in 41 languages.

Few cross-lingual transfer studies has been explored for the Indonesian language. However, prior studies are mostly investigated parsing or part-of-speech tagging tasks where there are already large corpora for Indonesian dataset in Universal Dependency Treebank[§] [1, 23, 33]. Ikhwantri [27] adopted a cross-character embedding between the English and Indonesian language and fine-tuned an English PLM to the NER task in Indonesian language. In our study, we compare several data transfer based and model transfer based models and analyze how the scenarios impact our low-resource settings of Indonesian NER. An interesting result from Rahimi et al. [53] demonstrated that Italian gives the best transfer to Indonesian in a direct model transfer method, compared to English as the most common single source language and Malay as the most similar language.

[§]https://universaldependencies.org/

3 Supervised Monolingual NER

This chapter explains the supervised method used for the monolingual NER for the Indonesian language. The subsections include; (1) Input representation for the models, (2) Bidirectional LSTM and CRF as the encoder-decoder model, and (3) Fine-tuning BERT-based models mentioned in Subsection 3.1 to downstream task such as NER. In these scenarios, we investigate the use of pre-trained language model in two different approaches, namely feature-based and fine-tuning. The former is expained in (1) as input representations for the (2) BiLSTM-CRF where the layers of the pre-trained models are frozen to obtain the vectors. The latter is explained in (3) where the parameters of the pre-trained models are fine-tuned to the small dataset of our NER task.

3.1 Input Representations

NER task dataset appears in a form of words with an entity label for each word. Feeding the words and labels directly to a neural network model would not work as neural networks understand vectors, not string characters. Therefore, an input representation is needed to convert the strings to vectors and capture underlying factors from the text data (Bengio, 2013). In this study, we use several input representation as the model variation to compare the performance of dynamic representation with the static input representation. We use FastText [8] as the static embedding as well as our baseline model. In terms of dynamic word embeddings, there are monolingual word embeddings, such as Flair [3] and IndoBERT [69], and multilingual word embeddings, such as multilingual BERT (mBERT) [15] and XLM-RoBERTa (XLM-R) [13]. The main difference between monolingual and multilingual word embeddings is the dataset adopted to pre-train the model. For monolingual, they use unlabeled corpora in one language,

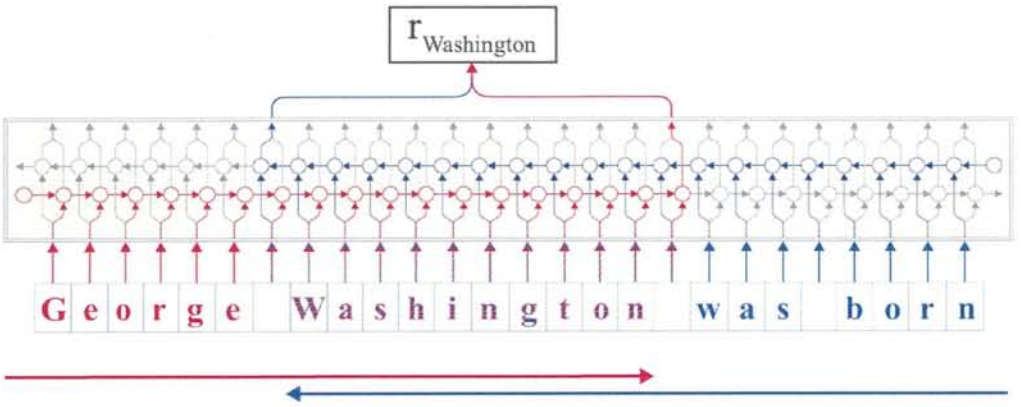

Figure 3.1: The architecture of Flair Embeddings [3].

while in the multilingual setting, they use unlabeled corpora consisted of text from many languages.

3.1.1 FastText

FastText is a word vector representation, or usually mentioned as word embeddings, that incorporates morphological information which is learned from the word's character n-gram [8]. FastText is pretrained in monolingual setting and available in many languages. It is known to overcome the OOV problems that commonly appear in low-resource settings since it allows the model to learn shared representation across words. We employ the use of FastText as the input representation for BiLSTM-CRF as our baseline model as implemented by [70].

3.1.2 Flair

Flair is a contextual string embedding pre-trained on a bidirectional LSTM backbone and it treats the words in a sentence as a sequence of characters. In this manner, the vector representation of each word has the underlying information about the context of the word used within the sentence. The representation of a same word in different sentences may vary since the preceding and the following characters of the word are not the same, depending on its contextual use. Therefore, it covers the strength of FastText [8] in solving the OOV problems since it

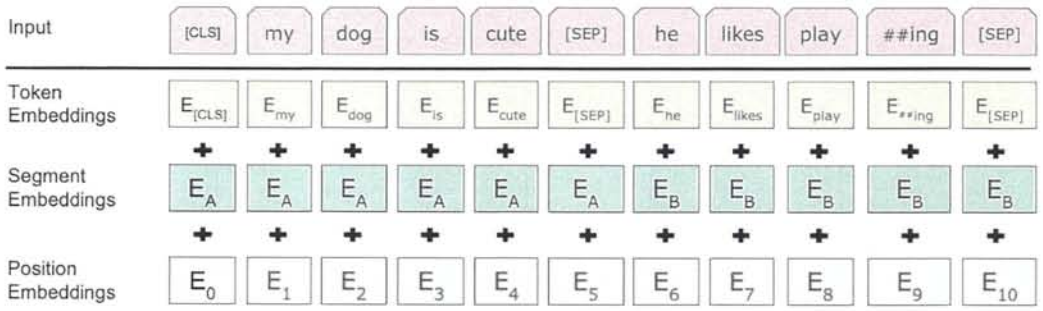

Figure 3.2: BERT input representation [15]. The token embeddings, segment embeddings, and position embeddings are summed to be its input embeddings.

has access to the character-level of the word, as well as understanding the context based on the surrounding text.

Figure 3.1 shows the architecture of Flair Embeddings when calculating the vector representation of each word. It consists of forward and backward networks, where the forward considers the sequence of characters from left to right, and the backward is from right to left. The vector representation of the word "Washington" is obtained by extracting the hidden state after the last character of the word for the forward network as shown in red. This hidden state incorporates the information all the way from the beginning of the sentence until the point of extraction. The same way is applied to the backward network as shown in blue, where the hidden state is extracted from the character before the first letter of the word and it contains the information from the end of the sentence, in reverse to the forward. The final contextual string embeddings are obtained by concatenating both hidden states.

3.1.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based pre-trained language model that consists of several transformer encoder and is pre-trained on huge unlabeled corpora, namely English Wikipedia (2,500 words) and BooksCorpus (800M words) [15]. In contrast to former pre-trained language models [3, 49], BERT does not have direction (left-to-right or right-to-left) and

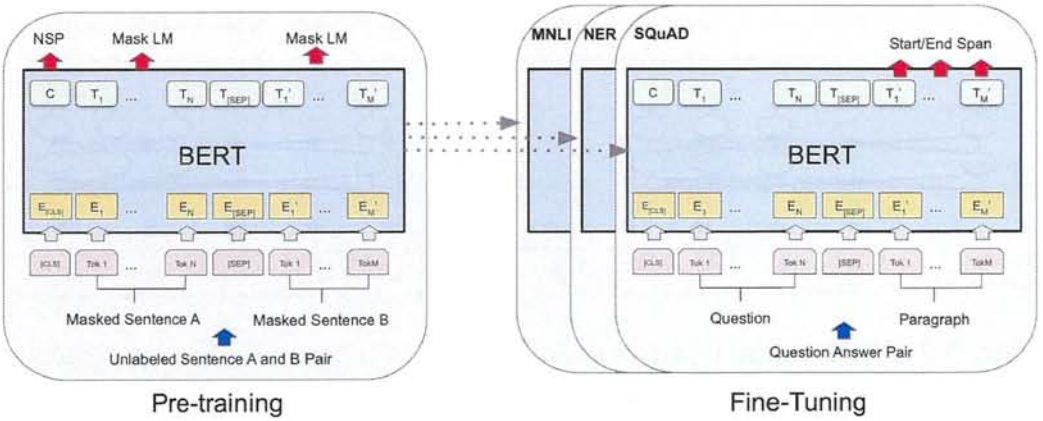

Figure 3.3: Pre-training and fine-tuning procedures for BERT [15].

learns the entire word sequence at once, based on its position embeddings. Figure 3.2 shows the architecture of BERT input embeddings. It sums token embeddings, segment embeddings, and position embeddings for its input representation. The position embeddings show the order of the words in a sequence of sentence, thus it does not need direction such as in LSTM.

BERT is pre-trained in two tasks to obtain the word representation in context, which are masked language modeling (MLM) and next sentence prediction (NSP). The MLM works in a way of cloze-task used in a second language learner exercise [61]. In this task, the BERT model learns to predict the randomly masked tokens with cross entropy loss. Meanwhile, the NSP learns the relationship between two sentences sequentially based on the segment embeddings as shown in Figure 3.2 and is aimed to support Question Answering (QA) and Natural Language Inference (NLI).

Devlin et al. [15] demonstrate that BERT performs well on many downstream tasks as well as NER and it can be implemented as an input representation for other models by freezing its layers as feature representation or fine-tuning the parameters to the downstream tasks as represented by Figure 3.3. Therefore, we applied both ways of exploiting BERT pre-trained model that elaborated in Sections 3.2 and 3.3 respectively. Most of BERT-based pre-trained models have two model sizes, the $\text{BERT}_{\text{BASE}}$ and the $\text{BERT}_{\text{LARGE}}$. The difference is the number of stacked encoder blocks used in the pre-training steps. The LARGE models usually

have twice as much as the BASE's encoders. However, for simplicity, all variant of BERT-based pre-trained models used in this study are the BASE models.

IndoBERT

IndoBERT is a variant of BERT model that was pre-trained on large Indonesian unlabeled corpora, Indo4B, which were gathered from widely available resource such as news, social media texts, and websites [69]. It is pre-trained using the same architecture of English BERT [15].

mBERT

Multilingual BERT (mBERT) is a variant of BERT model that was pre-trained on unlabeled corpora of 104 languages extracted from Wikipedia [15].

XLM-RoBERTa

XLM-RoBERTa is a multilingual version of RoBERTa, which is enhanced from the vanilla BERT model. The significant difference between BERT and RoBERTa is on the tasks used for their pre-traininig step and the masking pattern of the MLM task [43]. In contrast to BERT, RoBERTa removes the NSP task in the pre-training phase and only uses the MLM. Moreover, instead of using static masking, where the tokens are masked during the data processing, RoBERTa uses dynamic masking and duplicates the data ten times for the MLM task, in order that each sequence has ten different masking patterns. In the static masking, the masked tokens remain the same until the end of the pre-training, while dynamic masking changes the masked tokens every time a sequence fed in to the model. These strategies show better results that RoBERTa outperforms BERT in the downstream tasks.

XLM-RoBERTa was pre-trained on unlabeled corpora of 100 languages obtained from Wikipedia and CommonCrawl. In terms of dataset, XLM-RoBERTa was trained on a much larger size compared to mBERT, which is 2.5TB. Conneau et al. [13] show that XLM-RoBERTa achieves a good performance especially in low-resource languages.

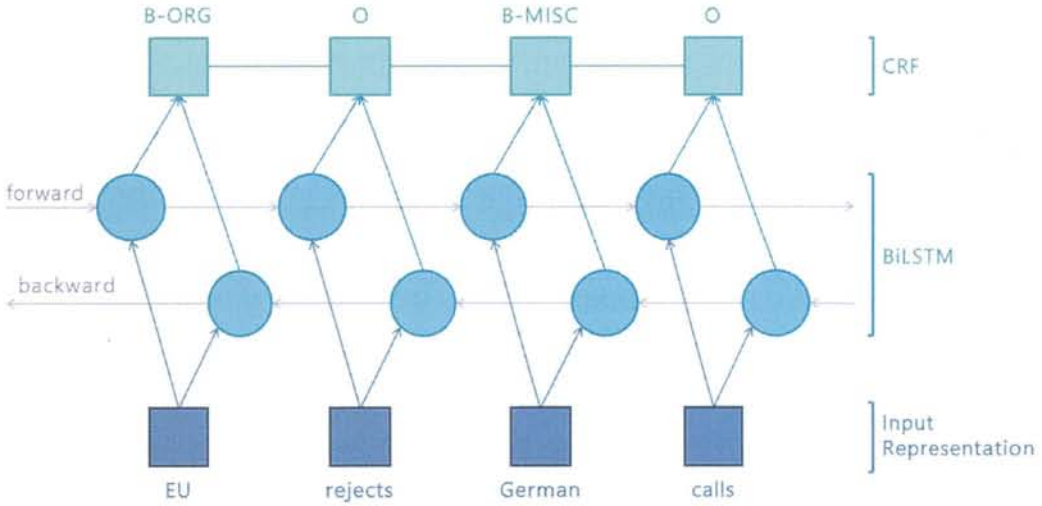

Figure 3.4: The architecture of BiLSTM-CRF [26].

3.2 Encoder-Decoder Model: BiLSTM-CRF

Bidirectional Long Short-Term Memory (BiLSTM) – Conditional Random Fields (CRF) was introduced by Huang et al. [26] and they obtained superior results for sequence tagging tasks such as POS tagging, chunking, and NER. They combined the advantages of BiLSTM that incorporates the past and future information of word sequence in a sentence and the CRF [36] with its ability to benefit the neighbor label information by focusing on sentence level knowledge, instead of token positions.

Figure 3.4 illustrates the architecture of BiLSTM-CRF sequence tagging model with its input representation as shown in the square boxes at the bottom. The input representations of features in time t could be one-hot-encoding for word feature, dense vector features, or sparse features. However, we used variety of input representation in this study as explained in Section 3.1. The input are fed into the BiLSTM networks that demonstrated in circles.

The BiLSTM utilizes both forward states and backward states for a specific time frame as sequence tagging task gives access to both past and future features respectively [21]. Each direction employs LSTM networks that are computed as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$
$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o)$$
$$h_t = o_t \tanh(c_t)$$

where i, f, o and c are the input gate, forget gate, output gate, and cell vectors, and σ is the logistic sigmoid function, with all are the same size as the hidden vector h. W is the weight matrix associated with each of its subscripts and the subscripts means as the name suggests. For example, W_{hi} is the hidden-input gate matrix and W_{xo} is the current input-output gate matrix. The forward and backward networks are implemented for the whole sentences and the hidden states is reset to 0 every time the networks learn a new sentence. The hidden states from forward and backward networks obtained from the BiLSTM are given to the CRF layer to predict the output labels.

A CRF layer can efficiently benefit the past and future labels to predict the current label by using its state transition matrix as parameters. This is similar to the BiLSTM networks in employing the past and future input features, so that they both manage to complement each other in exploiting information of the input representations and labels in a sequence tagging task.

All experiments of the BiLSTM-CRF in this study is trained using FlairNLP*, an NLP toolkit for sequence and document classification task by Humbolt University of Berlin [2].

3.3 Fine-tuning

Fine-tuning is a way of taking advantages from BERT-based model by fed in task-specific input to the model without major task specific architecture modification [15]. This approach allows all BERT-based models' parameters to adjust end-to-end with the input from a downstream task and obtain the output by

*https://github.com/flairNLP/flair

applying one classification layer fit in the type of the task, whether it is a single sentence classification, token classification, or sentence pair classification tasks. Fine-tuning approach is relatively less expensive compared to pre-training step without lessening the performance of the downstream tasks. Regarding the use of fine-tuning approach in this study, we only fine-tune the BERT-based pre-trained models, namely IndoBERT [69], mBERT [15], and XLM-RoBERTa [13].

4 Unsupervised Cross-Lingual NER

This chapter explains the unsupervised study of cross-lingual knowledge transfer from other languages to the Indonesian language. The subsections include two different method used in the transfer learning, namely Data Transfer and Model Transfer. The detail of source languages and datasets are explained in the next chapter. We illustrated a fully unsupervised learning of the Indonesian NER model that all of the training data were obtained by transferring the knowledge from other languages and only used the test set of our gold Indonesian NER dataset to evaluate the models.

4.1 Data Transfer

The idea of the data transfer is translating the NER dataset from other languages to the Indonesian language and make use of the translated pseudo-data as the training data for the Indonesian NER model. Due to the limited size of our gold NER dataset in the Indonesian language, we want to take benefit from the larger labeled data of other languages that are available publicly.

4.1.1 Vector-based Transfer

Vector-based transfer is a method in machine translation that translates sentences in an unsupervised way and highly depends on the source and the target languages' monolingual word embeddings by projecting them in a shared embedding space [38, 39]. By aligning the vectors from both word embeddings, two words in different languages can be taken as a word-pair based on the nearest neighbor. Figure 4.1 illustrates the vector projection in translating the two languages.

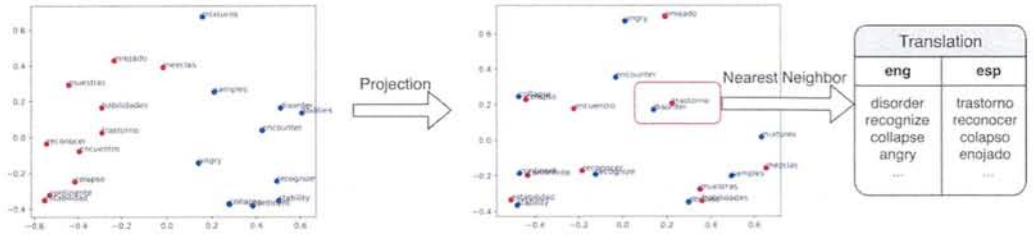

Figure 4.1: Example of the source-target languages' vector projection in a shared embedding space [75].

We translated the source languages' NER dataset to the Indonesian language and directly copied the labels without changing the order using any word alignment tool, considering that the vector-based transfer performs a word-to-word translation of the language pair [75]. We used a vector-based translation tool MUSE* developed by Facebook Research [38, 39]. After obtaining the translated data in the Indonesian language along with the labels of the source languages' dataset, we trained the NER model using the translated pseudo-data with FlairNLP† [2], the same tool we used in all experiment for BiLSTM-CRF model in this study.

4.1.2 Neural Machine Translation (NMT)

For comparison, we also performed the translation using NMT which highly depends on parallel data and NMT model [56, 67]. We trained an NMT model to translate the NER dataset from the source languages into the Indonesian language, and instead of mapping the labels using attention as opposed in previous works, we project the labels using a word alignment tool. In training the NMT model, we use Fairseq toolkit‡ [48] with default hyper-parameters of the Transformer-based model implementation. To align the words from the source to the translated target language for projecting the labels, we use a word alignment tool Eflomal§ [79] that has better performance than its predecessors. The overall

*https://github.com/facebookresearch/MUSE
†https://github.com/flairNLP/flair
‡https://github.com/pytorch/fairseq
§https://github.com/robertostling/eflomal

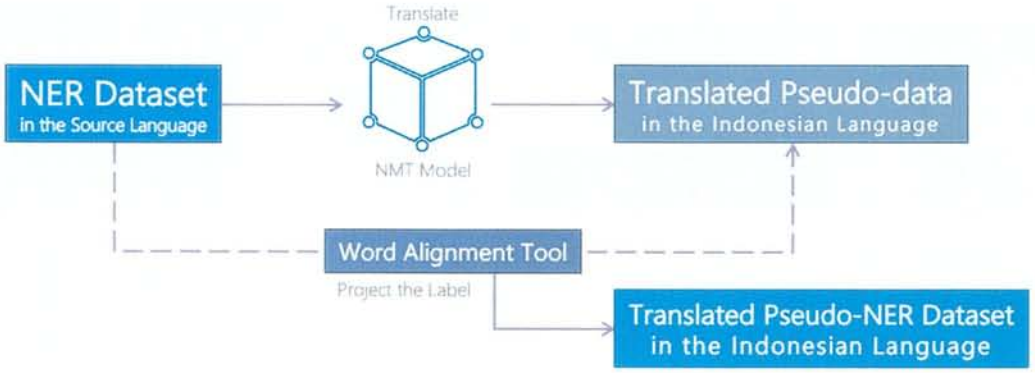

Figure 4.2: The overview of our NMT Transfer process to create the translate pseudo NER dataset in the Indonesian language.

process of our NMT transfer process is shown in Figure 4.2.

4.1.3 Parallel Data

To exploit the high availability of parallel English-Indonesian corpora, we also demonstrated a transfer method using parallel corpora. Instead of translating an NER dataset of the source language, we annotate the source side of the parallel data using Stanza[¶] [52], a state-of-the-art syntactic analysis toolkit for many languages by Stanford NLP. Once we obtained the annotations for the source side, we then project the annotations to the target side of the parallel data, which is the Indonesian language, since the Stanza is not available for the Indonesian language. We also used Eflomal[‖] [79] to project the labels from the source to the target side. Figure 4.3 presents the overall process of pseudo NER dataset creation from parallel corpora.

4.2 Model Transfer: Teacher-Student Learning

In this section, we examine the cross-lingual transfer method that highly depends on the availability of labeled data in the source language and language-

[¶]https://stanfordnlp.github.io/stanza/
[‖]https://github.com/robertostling/eflomal

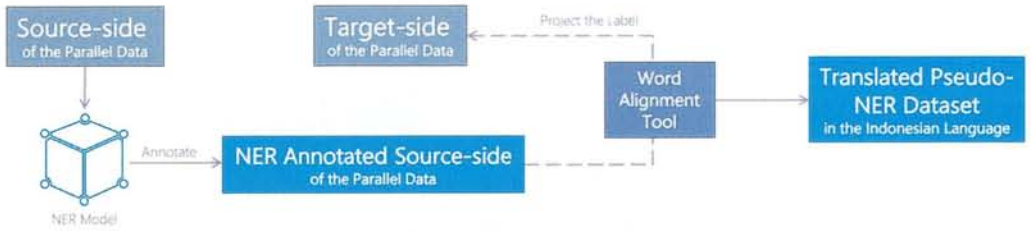

Figure 4.3: The overview of our data transfer process from parallel corpora to create the translate pseudo NER dataset in the Indonesian language.

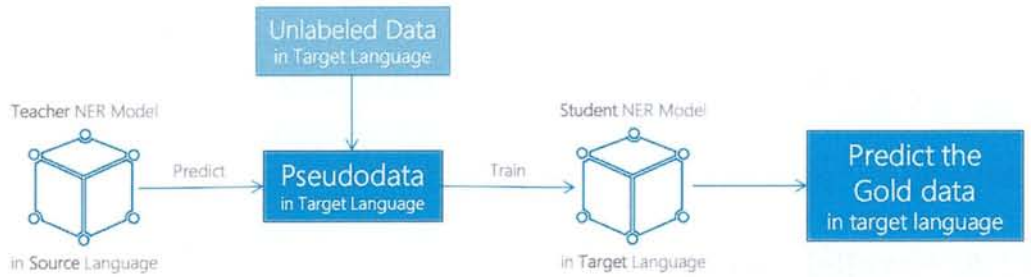

Figure 4.4: The teacher-student learning from Teacher model in the source language to Student NER model in the Indonesian language.

independent features models, such as cross-lingual word representations [13, 15]. Former studies of model transfer method in cross-lingual settings by focusing on training a shared NER model on the source languages' labeled data and testing the model directly on the target language [72, 73]. In the teacher-student model [71], we first train a teacher model on the source language NER dataset and this teacher model is used to predict NER labels for the target language. Up to this point, the teacher model works similarly to directly model transfer. However, the labeled dataset obtained from the teacher model's prediction is adopted as a pseudo-labeled data to train a student model in the target language, which produce the final results of the NER model prediction. The general steps of teacher-student model are illustrated in Figure 4.4.

5 Datasets and Experiments

This chapter defines the datasets and the experiment scenarios executed for the supervised monolingual and unsupervised cross-lingual transfer learning as well as the dataset inconsistency and re-annotation of the existing available Indonesian NER dataset.

5.1 Monolingual Indonesian NER

In this section, we explain the inconsistency that appears in the existing Indonesian NER dataset by S&N (2016) [59] and the way we conducted the re-annotation step as well as the annotation guideline that is used to enhance the dataset.

5.1.1 Inconsistency of the Current Indonesian NER Dataset

We examined an available open dataset by S&N (2016)*, yet we found some problems such as entities that are not tagged as it is or non-entities that are tagged as entities. We show the inconsistent example of person and organization entities of the annotation in Table 5.1. The three sentences include a same pattern, [title][organization][person]. However, in the first sentence, all the tokens *"Ketua Umum Gerindra Prabowo Subianto"* as a person's name. In fact, those tokens are three different entities. In the second sentence, red tokens indicate inconsistent annotation, where the token *"Suhardi"* (a person's name) is the only token that is labeled as an entity by S&N (2016), without labeling *"Partai Gerindra"* as an organization. On the other hand, the blue tokens present the correct annotation. In the second sentence, the token *"Politikus"* is not labeled as an entity, but *"PDI*

*https://github.com/yusufsyaifudin/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	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 5.1: 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.

Perjuangan" and *"Guruh Soekarnoputra"* are labeled as organization and person, respectively. So as the third sentence, *"Wakil Ketua Umum"* [title], *"Partai Gerindra"* [organization], and *"Edy Prabowo"* [person]. These examples illustrate distinct annotations in the same pattern. We noticed this occurs several times, so we checked through the annotations in the dataset one by one.

5.1.2 Data Re-annotation

The re-annotation process was manually done by three native speakers. Despite the five entities included in S&N (2016), we re-annotated three common entities in NER: location, organization, and person. We exclude the time and quantity entities since our focus is a model that can recognize ambiguous entities in the Indonesian language, such as organization and person. Time and quantity are often written in numeric form, so a robust NER model will readily distinguish them. Therefore, we only compared the three entities in both datasets to make the results fairly comparable.

In terms of dataset split, we use the same test set as in S&N (2016) and randomly sampled some instances from the training set to form a development set, as presented in Table 5.2a. We follow the BIO format by Tjong Kim Sang

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

(a) Data statistics.

		Our Tags				# of
		LOC	ORG	PER	O	tags
S&N 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	

(b) Confusion matrix of our re-annotation from S&N (2016).

Table 5.2: Data Statistics and 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.

et al. [63] as the standard NER dataset format. The subsection below describes our annotation guidelines used in the re-annotation process.

5.1.3 Annotation Guideline

In this part, we include our guidelines in re-annotating the dataset to define each entity clearly. This guideline is made by native speakers, both based on the characteristics of the Indonesian language [10] as well as the intuition as native speakers (e.g., Location and Person names have clear characteristics to be categorized as such. In terms of organization, it is often the ambiguous one, so we checked if the name exists as an organization or not). We also use the English NER entity labels as a reference, although it's not directly the same because the Indonesian language has some differences in language features and common entities that are used.

- `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* $[Ical]_{B\text{-}PER}$*"*, not as *"*$[Pak]_{B\text{-}PER}$ $[Ical]_{I\text{-}PER}$*"*.

- 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:

  $[\text{Universitas}]_{B\text{-}ORG}$ $[\text{Gadjah}]_{\text{I-ORG}}$ $[\text{Mada}]_{I\text{-}ORG}$ $([\text{UGM}]_{B\text{-}ORG})$ berlokasi di $[\text{Yogyakarta}]_{B\text{-}LOC}$

Regarding ambiguous entities, we determined the tags following the word's semantic meaning. For example, the organization name tends to confuse with the location name, either because of the preposition or the dual meaning of the name as an organization or the organization's office where activity happens. We use Fleiss' kappa [18] to calculate the inter-annotator agreement of the three annotators and got a score of 0.92, which shows good reliability [5].

We present the label changes on our annotation in Table 5.2b. The number of location entities reduces approximately 20%, and almost 500 tokens of non-entities decreased by about 500. In contrast, the organization and person entities increased after the re-annotation. As shown in the table, S&N (2016) did not

correctly label most of the organization entity, where 701 organization tokens were labeled as non-entities.

5.2 Cross-lingual Transfer Learning

We investigate two approaches in conducting the unsupervised cross-lingual transfer learning for the Indonesian NER with the methods explained in Chapter 4. The first approach is single-source transfer, with English as the source languages due to its largest available corpora in many aspects, including both labeled NER and parallel dataset. The second approach is multi-source transfer, since we want to broaden the scope of the source languages in examining cross-lingual transfer. Table 5.3 summarizes the data statistics of the source languages' NER dataset.

Although the Indonesian language does not belong to a language family that in the category of high-resource language in the NLP field, we chose the languages with the most available NER dataset after English. For this reason, we include English, Spanish, Dutch, and German as the source languages in the multi-source approach. All datasets have more than the three entity types we use (person, location, and organization) so that we omitted the entity types other than these three to perform comparable transfer learning to our Indonesian NER dataset.

All of the training data for the Indonesian NER models in these approaches are using the pseudo-labeled data obtained by transferring the knowledge from the source languages' dataset and only used our gold Indonesian NER development and test dataset to evaluate the models. In the model transfer method (Section 4.2), we omit the labels from our gold Indonesian NER train dataset when creating the pseudo-labeled data from the Teacher model.

5.2.1 Single-source

For the English NER source dataset, we use the CoNLL-2003 Dataset [65]. This dataset is used in the vector-based transfer (Subsection 4.1.1), the NMT transfer (Subsection 4.1.2), and the model transfer (Section 4.2). We also extracted the Indonesian Wikipedia and took about 30K sentences randomly from the dump. We used this unlabeled dataset as additional training data for the model transfer-based method, since the teacher model of the model transfer method predicts an

Data Split	English (EN)		Spanish (ES)		Dutch (NL)		German (DE)	
	Sentence	Entity	Sentence	Entity	Sentence	Entity	Sentence	Entity
Train	14,987	23,499	8,323	18,798	15,806	13,344	24,000	21,215
Development	3,466	5,942	1,915	4,351	2,895	2,616	2,200	1,790
Test	3,684	5,648	1,517	3,558	5,195	3,941	5,100	4,495
Total	22,137	35,089	11,755	26,707	23,896	19,901	31,300	27,500

Table 5.3: Data statistics of our source languages' NER Datasets. The English is from CoNLL-2003 [65], the Spanish and Dutch are from CoNLL-2002 [64], and the German is from GermEval 2014 [7].

unlabeled dataset from the target language to be pseudo-training data for the student language. The objective is to compare if using a larger noisy dataset from Wikipedia would improve the model's performance.

In terms of data transfer method using parallel corpora (Section 4.1.3), we did not use any NER dataset of the source language since we annotated the source-side of the parallel corpora with Stanza. The parallel English-Indonesian corpora we use are Asian Language Treebank (ALT) [55], Global Voices v2018 [62], BPPT Parallel Dataset, and SMERU, AusAid, and BBC from the ID-EN Bilingual Corpus[†]. Due to the absence of entities in some sentences that makes the pseudo-data becomes more noisy, we removed the sentences that do not have any entity.

Regarding the NMT model we employed to translate the English NER dataset to the Indonesian language, we use the IWSLT 2016 TedTalks dataset [11].

5.2.2 Multi-source

We adopted English (EN), Spanish (ES), Dutch (NL), and German (DE) for the multi-source approaches. The English dataset is CoNLL-2003 [65], the Spanish and Dutch NER datasets are from CoNLL-2002 benchmark [64], and GermEval 2014 [7] is used as the German NER dataset. For the multi-source approach, we only implement the vector-based method (Subsection 4.1.1) and model transfer

[†]https://github.com/desmond86/Indonesian-English-Bilingual-Corpus

(Section 4.2 due to the limited available parallel data for Spanish, Dutch, and German with the Indonesian language.

For the model transfer, we implement the BiLSTM-CRF approach with XLM-RoBERTa [13] instead of fine-tuning the pre-trained language model such in Wu et al. [71]. Our study shows that the BiLSTM-CRF approach performed generally better than fine-tuning the transformer-based pre-trained model. We chose to use XLM-RoBERTa since the IndoBERT does not cover characters other than basic latin as used in the Indonesian language. Meanwhile the XLM-RoBERTa was pre-trained on many languages including Spanish and German which have the Latin-1 supplement. When creating the pseudo-data using the teacher models in multi-source, we follow the majority voting scheme [51, 71].

We have five scenarios for the multi-source approach, (1) **EN-ES-NL-DE**: all of the source language, and the rest is an ablation study with one-left-out manner; (2) **EN-ES-NL**, (3) **EN-ES-DE**, (4) **EN-NL-DE**, and (5) **ES-NL-DE**.

5.3 Experiment Settings

All models are evaluated using the standard accuracy measurement for the NER task, which is F_1-score, along with the precision and recall.

BiLSTM-CRF. We adopt the implementation of BiLSTM-CRF by FlairNLP FlairNLP[‡], an NLP toolkit for sequence and document classification task by Humbolt University of Berlin [2]. We experimented five times for each model and averaged the scores to ensure the consistency of the models. We also include the standard deviation of the the overall F_1-score to show the amount of variation of the scores. The dataset format follows IOB format by Tjong Kim Sang [63] with three entity types, which are location (LOC), organization (ORG), and person (PER). The first approach with BiLSTM-CRF is done by using different input representations, namely FastText [20], Flair [3], mBERT [15], XLM-R [13], and IndoBERT [69], with parameter settings as follows: a mini-batch size of 32, one BiLSTM hidden layer, 256 BiLSTM hidden units, a dropout of 0.5, and a learning rate of 0.1. The framework implements an early stopping method, so we set a

[‡]`https://github.com/flairNLP/flair`

maximum number of 200 epochs and it will stop training as the model converges.

Fine-tuning BERT-based models. We fine-tune the BERT-based PLMs; mBERT and XLM-RoBERTa and we set the batch size to 32, the learning rate to [1e-5, 3e-5, 5e-5, 7e-5], and 5 epochs for each model. We ran the experiment of each model for five times with different seed to average the scores and calculate the standard deviation as similarly done for the BiLSTM-CRF method.

6 Results and Analysis

Tables 6.1, 6.2, 6.3, and 6.4 summarize our experimental results. Our re-annotation exhibited superior performance compared to the annotation of S&N (2016) when trained using the baseline model. Our experiments comparing the monolingual and multilingual word embedding yielded positive results when using the IndoBERT as feature representation for the BiLSTM-CRF architecture. The cross-lingual setting with word vector shared embedding representation also shows a competitive result to our baseline in the monolingual scenario.

6.1 Results

6.1.1 Annotation performance

We present the comparison of both annotation performances in Table 6.1. To investigate both models' performance in the same setting, we did a cross-test for each model by testing them on both test sets, as shown in the table. In the baseline model, testing the S&N (2016) on both test sets shows different scores. By testing on the S&N (2016), we get an F1 score of 76.11 and on Ours, 84.41, especially for the problematic organization tag, of which the score jumped by about 20 points. The significant score differences of organization entity happened on both test scenarios, demonstrating that the occurrence of organization tokens that was not tagged as shown in Table 5.1 led to a sharp drop of S&N (2016)'s performance. Testing with Ours presents more even scores between the entities and a relatively high overall F1 score at 90.85. This demonstrates that the inconsistency in the dataset could cause a low prediction score, and our re-annotation improved the model performance.

Dataset Annotation		Overall Scores			LOC			ORG			PER		
Train	**Test**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
S&N	S&N	69.20	84.55	76.11	74.82	85.78	79.92	45.54	83.02	58.82	83.60	84.86	84.23
Ours	S&N	66.39	88.28	75.79	79.05	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	**92.23**	**89.52**	**90.85**	**89.02**	**88.52**	**88.76**	**89.13**	**88.13**	**88.63**	**95.50**	**90.83**	**93.10**

Table 6.1: 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.

6.1.2 Monolingual vs multilingual pre-trained models

We present the results of using pre-trained monolingual and multilingual BERT models for our Indonesian NER task in Table 6.2. The use of IndoBERT pre-trained model as a feature representation for BiLSTM-CRF architecture shows the best score at 94.90. Both ways of exploiting IndoBERT, as feature representation and fine-tuning, yield very high organization scores. The IndoBERT used the `Indo4B` dataset in their pre-training step, which contains Indonesian news corpora, the same source as our NER dataset [69]. Therefore, the rich Indonesian vocabularies covered by IndoBERT, the domain similarity between the `Indo4B` and our dataset, and BiLSTM-CRF with its sequence-based architecture fits the Indonesian NER task better.

In the multilingual settings, the XLM-R performs better compared to the mBERT. As richer and larger dataset gives more learning to a deep learning model, so as in XLM-R, where it is pre-trained using a large-scale unsupervised multilingual data compared to mBERT [13, 15]. The XLM-R model could work better in an NER task because entity names, particularly organization, sometimes originated from English or other languages [69]. However, most of the entity names in our dataset are in the Indonesian language, thus IndoBERT well suits our vocabularies. Furthermore, mBERT and XLM-R applied similar tokenization for their words where it splits a longer token into more common subwords [32, 74]. When freezing the word representation of the pre-trained model for the input of BiLSTM-CRF, we averaged the vector of the subwords. The final average score

Models	Features	Overall Scores			F_1 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
	Flair	89.75	91.05	90.40±0.04	90.64	86.70	93.32
	mBERT	90.11	89.80	89.95±0.05	87.30	86.54	93.20
	XLM-R	91.12	94.05	92.56±0.54	89.01	88.95	96.25
	IndoBERT	**94.93**	**94.87**	**94.90±0.31**	**89.84**	**92.99**	**97.49**
Fine-tune mBERT	–	89.54	92.33	90.91±0.51	**86.73**	87.78	94.11
Fine-tune XLM-R	–	91.60	**94.76**	**93.15±0.23**	85.59	90.46	**97.41**
Fine-tune IndoBERT	–	**91.66**	94.64	93.13±0.31	83.84	**91.81**	96.83

Table 6.2: Supervised monolingual Indonesian NER model performance for contextual embedding experiment. The multilingual pre-trained models are mBERT and XLM-R; others are monolingual.

of each word is inevitably biased since multilingual models are trained on many languages, so that the sub-word representations are shared from other languages as well.

6.1.3 Cross-lingual transfer learning

The cross-lingual transfer learning experiment results are divided into two parts following the approaches explained in Section 5.2, namely single-source and multi-source.

Single-source

Table 6.3 shows the results of the single-source cross-lingual transfer from English as the source language to Indonesian as the target language. Rows 1-9 show the data transfer based results, with vector-based transfer on rows 1-3, NMT on rows 4-6, and parallel data on rows 7-9. Both NMT and parallel data used Eflomal to align the word translations. The last four rows are the Teacher-student model, comparing the use of our IDNER dataset and Indonesian Wikipedia dumps as the unlabeled training data and XLM-RoBERTa and multilingual BERT as the transformer pre-trained models.

Method	Training Data	Model	P	R	F
Vector-based Transfer	CoNLL2003 English NER	BiLSTM-CRF (with IndoBERT)	**90.96**	**87.80**	**89.35+0.44**
		Fine-tune XLM-RoBERTa	79.19	86.01	82.45±0.33
		Fine-tune mBERT	71.64	78.49	74.91⊥0.70
NMT-based with Effomal		BiLSTM-CRF (with IndoBERT)	88.62	85.02	86.78±0.51
		Fine-tune XLM-RoBERTa	68.58	69.42	68.99±2.09
		Fine-tune mBERT	69.59	69.80	69.69⊥1.01
Parallel Data with Effomal	Global News Parallel Data	BiLSTM-CRF (with IndoBERT)	84.68	80.02	82.28±0.76
		Fine-tune XLM-RoBERTa	77.60	83.05	80.23±1.14
		Fine-tune mBERT	72.17	78.39	75.15±0.68
Teacher-Student Learning	IDNER-News-2K	Fine-tune XLM-RoBERTa	79.56	87.73	83.44±0.51
		Fine-tune mBERT	70.81	80.73	75.45±0.72
	IDWiki-dumps	Fine-tune XLM-RoBERTa	78.67	87.04	82.65⊥0.69
		Fine-tune mBERT	70.54	81.98	75.83±0.66

Table 6.3: Single-source cross-lingual transfer results comparing the data transfer and the model transfer-based methods.

BiLSTM-CRF with IndoBERT always obtains the best performance for each data transfer scenario. Translating a gold NER dataset from the source language using vector-based transfer achieved the best score. This happened because the vector-based transfer performs a word-to-word translation where the word order is not changed like in NMT or parallel data. Meanwhile, when comparing the performance of using NMT or parallel data to which we projected the labels based on their word alignment, Table 6.3 shows that the fine-tuning works better on the parallel data and BiLSTM-CRF works better on the translated data from the CoNLL-03 English NER data.

Multi-sources

We summarize the results of the multi-source cross-lingual transfer experiment in Table 6.4. Overall, the vector-based transfer outperforms the teacher-student learning method with Spanish, Dutch, and German as the source languages. In the vector-based transfer scenario, Dutch gives the most gain to be transferred to the Indonesian language since both nations have a long historical background, as well as almost 6,000 Indonesian words are borrowed from the Dutch [57]. The vector-based transfer method highly depends on the monolingual models of source and target language. The monolingual word representations used to perform the

Method	Languages	Precision	Recall	F_1 Score
Vector-based Transfer	EN (single-source)	88.40	86.47	87.42±0.70
	EN-ES-NL-DE	89.19	86.83	88.00±0.43
	EN-ES-NL	88.08	**89.46**	88.76±0.64
	EN-ES-DE	89.78	86.69	88.21±0.29
	EN-NL-DE	89.65	87.42	88.52±0.28
	ES-NL-DE	**90.70**	87.69	**89.17±0.43**
Teacher-Student Learning	EN (single-source)	83.76	88.50	86.07±0.52
	EN-ES-NL-DE	85.09	86.66	85.86±0.18
	EN-ES-NL	82.79	86.22	84.47±0.27
	EN-ES-DE	85.14	86.66	85.89±0.50
	EN-NL-DE	**85.32**	**87.26**	**86.28±0.29**
	ES-NL-DE	84.84	85.00	84.92±0.27

Table 6.4: Multi-source cross-lingual transfer results from English (EN), Spanish (ES), Dutch (NL), and German (DE) as the source languages. All models were trained on BiLSTM-CRF with XLM-RoBERTa for the input representation.

vector-based transfer are FastText-based, where the word vectors are built based on the words' internal structure and morphology. The Indonesian language has similar morphology patterns to the Dutch and also use the same Latin alphabet. The consonant-vowel-consonant-vowel patterns significantly appear in a word, compared to German and English that the words often have three or more consecutive consonants. Therefore, similarity in the words' morphology benefits the cross-lingual transfer using the vector-based method.

Although the teacher-student learning did not perform as well as the vector-based, it still has a competitive result that is lower about three points from the vector-based method. However, in opposition to the vector-based transfer, the German gives the most significant gain. It is shown that the lowest score in the teacher-student learning results is the model without the German NER dataset as the source (EN-ES-NL). We hypothesize that this happened because the largest number of sentences comes from German, about 31K, where the other language sources have less than 20K sentences. Considering the model transfer relies on

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 6.5: Examples of errors in prediction comparing S&N (2016) and our annotation when trained using baseline BiLSTM-CRF model. The tokens in red indicate incorrect predictions, and those in blue indicate the correct ones.

the cross-lingual representation, the XLM-RoBERTa has English, Indonesian, and German as the highest amount of data used in the pre-training steps among the source languages we used in this experiment.

In general, both vector-based and model transfer methods in the multi-source scenario improve the performance of cross-lingual transfer learning compared to the single-source scenario. Using multiple languages as the source transfer demonstrated a practical approach for the Indonesian NER task without ignoring the transfer methods used and the source-target language's similarity.

6.2 Discussion and Analysis

This section presents some error examples of the model trained on S&N (2016) annotation compared to our annotation. Moreover, we clarify the effect of monolingual and multilingual information on the prediction result.

Label		BiLSTM-CRF				Fine-tune			Total
Gold	Predicted	FastText	mBERT	XLM-R	IndoBERT	mBERT	XLM-R	IndoBERT	
LOC	PER	4	4	4	4	3	4	4	27
	O	6	9	1	5	1	6	5	33
	ORG	0	0	0	2	0	0	3	5
O	LOC	5	8	10	5	12	6	11	57
	ORG	28	64	64	25	28	26	28	263
	PER	8	15	14	7	15	4	6	69
ORG	LOC	2	7	9	1	7	4	3	33
	O	38	36	16	18	31	20	16	175
	PER	1	6	0	0	7	0	0	14
PER	O	33	17	2	4	10	4	4	74
Not prefixed with B-		0	0	0	1	19	23	12	55
Total		125	166	120	72	133	97	92	

Table 6.6: The number of errors in prediction results of each model using our re-annotated dataset that are presented on Table 6.2.

6.2.1 Re-annotation

Sentence 1 of Table 6.5 demonstrates the annotation errors of S&N (2016) and its prediction in the case of recognizing a person's name. The words *"Ketua Umum Gerindra"* were labeled as part of a person's name, whereas it is a title of a person. When predicting the tokens, the S&N (2016) model correctly spots *"Prabowo Subianto"* as a person's name, but it missed the *"Gerindra"* as an organization name. On the other hand, our model appropriately does not recognize *"Ketua Umum"* as an entity and both *"Gerindra"* and *"Prabowo Subianto"* as they are.

In Sentence 2, the S&N (2016) annotation did not tag the words *"Kementerian Hukum dan Hak Asasi Manusia"* ("The Ministry of Law and Human Rights" in English) as an organization name. However, both models predicted all of the tokens accurately. These errors exhibit how inconsistency in a labeled dataset impacts the inference of a model and worsens the model's score. Both sentences resulted in false-positive cases in the evaluation step. When the model prediction was correct but the annotation was incorrect, the prediction was considered as a false prediction.

6.2.2 Model prediction in monolingual and multilingual settings

To investigate the impact of monolingual and multilingual pre-trained models, we performed a thorough analysis of the prediction resulted from each model in Table 6.2, particularly in the organization entity case. Table 6.6 presents the number of errors that appeared on each model. During the error analysis, we only counted the errors when the model falsely predicted the entity and ignored the prefix difference of the labels unless it did not begin with "B-". Aligned with our findings during the re-annotation, the organization entity has the highest number of errors, both when the non-entities are recognized as the organization and vice versa. The second problem happened on the person entity, where the model falsely predicted non-person names. In the case of FastText model, it has a moderate number of errors in Organization and Person entities. They are mostly because of the OOV problem, where the tokens were not present in the train and development sets.

Most of the errors on the O→ORG case in the IndoBERT (25 cases when used with BiLSTM-CRF and 28 cases when fine-tuned) happened because the models are too context-sensitive so that it recognized tokens that have an underlying meaning of an organization as an organization entity (e.g., tokens started with the word "lembaga" (institution) or "gerakan" (movement), nevertheless these tokens did not mention any official organization name). Meanwhile, the multilingual models' errors happen because the tokens are prefixed with capital letters (Xxxxx) or have all capital letters form (XXX). This phenomenon illustrates that the multilingual models are great at learning the internal word structure of a language. In the case of ORG→O, most errors occurred in mBERT, because the organization tokens are not started with a capital letter, so does in the case of XLM-R. While this type of error in IndoBERT happened because of the OOV problem.

Regarding O→PER, it appeared because of a similar reason with O→ORG in IndoBERT. The model falsely recognized the tokens that mentioned a title of a person or referred to a person. It contextually recognized the tokens as a position of a person name, but the token did not mention one. The last type is when the model predicted entities without the prefix "B-" for the first token. It

Sentence 1	The Nasdem Party is different to Nasdem as a societal organization					
Indonesian	**Partai**	**Nasdem**	**berbeda**	**dengan**	**ormas**	**Nasdem**
English Translation	Party	Nasdem	different	to	societal-organization	Nasdem
Our annotation	B-ORG	I-ORG	O	O	O	B-ORG
BiLSTM-CRF + IndoBERT	B-ORG	I-ORG	O	O	B-ORG	I-ORG
Fine-tune XLM-R	B-ORG	I-ORG	O	O	O	I-ORG

Sentence 2	The head of regional election through the regional people's representative council							
Indonesian	**pemilihan**	**kepala**	**daerah**	**melalui**	**dewan**	**perwakilan**	**rakyat**	**daerah**
English Translation	election	head	regional	through	council	representative	people	regional
Our annotation	O	O	O	O	B-ORG	I-ORG	I-ORG	I-ORG
BiLSTM-CRF + IndoBERT	O	O	O	O	B-ORG	I-ORG	I-ORG	I-ORG
Fine-tune XLM-R	O	O	O	O	O	O	O	O

Table 6.7: Error examples from the contextual embedding experiments in monolingual (IndoBERT) and multilingual (XLM-R) settings. The tokens in red indicate incorrect and those in blue indicate the correct one.

occurred mainly in the fine-tuned model, where it does not have the CRF as the encoder model. We hypothesize that it happened because the fine-tuning only uses the softmax function for the classification layer, where it decides the final output based on probability score, without considering the previous tag as in CRF. Therefore, in some instances, it is possible that the model lost some critical information, such as the beginning or the middle of the entity.

Table 6.7 presents some error examples mentioned in Table 6.6. The first sentence illustrates when the use of IndoBERT for input representation for BiLSTM-CRF made the model too sensitive to the context. The sentence mentions two "Nasdem": Nasdem as a party and the Nasdem as a societal organization. The name of Nasdem as a Party includes the word "Partai" (party) in its legal name; meanwhile, the Nasdem as a societal organization does not include the word "ormas" (societal organization). In this case, the BiLSTM-CRF + IndoBERT also recognized the word "ormas" as a part of the organization name since it refers to an organization, without looking at its word shape (Xxx or xxx).

On the other hand, the fine-tune XLM-R correctly does not include the word "ormas" but falsely recognized the "Nasdem" as I-ORG, not B-ORG. Aligned with our argument for Table 6.6 about fine-tuning, it determines the final output based on the highest probability without looking at the previous tokens, so because "Nasdem" is usually preceded by another word–in this case "Partai"–the model

is more likely to recognize it as I-ORG. The second sentence demonstrates that the fine-tune XLM-R relies on the morphological feature. Therefore, when an organization name does not start with a capital letter (Xxx), it does not recognize it as an entity name.

7 Conclusion

We built a more consistent Indonesian NER dataset by re-annotating a previously inconsistent dataset and made it available publicly for further use to the research community*. Our annotation resulted in an F_1 score of 90.85 with the baseline, FastText as the input representation. We also compare the use of monolingual and multilingual BERT-based pre-trained models to obtain a more robust model in tackling word ambiguity problems in the Indonesian NER task. We found that the sequence model architecture of BiLSTM-CRF combined with the monolingual IndoBERT pre-trained model yielded a very high F_1 score of 94.90.

The monolingual and multilingual models have different error causes during the inference. The monolingual model, IndoBERT, is highly context-sensitive to our organization entity problem, so that most of the errors occurred because of its sensitiveness to the context of the words. On the other hand, the multilingual models might have less common vocabulary to Indonesian organization names, so they depend on the morphology of the words to recognize the entity names.

Finally, we show that both single-source and multi-source cross-lingual transfer learning from the high-resource languages give a very competitive result using the data transfer method by projecting the entity label using vector-based word-to-word translation. Interestingly, we found that Dutch provides competitive results compared to English, as the language with the highest resource available, due to shared vocabularies between Dutch and Indonesian, both morphologically and phonetically. As our result from the cross-lingual transfer experiments, we argue that multilingual transfer learning can be an alternative to the low-resource languages that may be done at a lesser cost.

For future works, we are interested in investigating each language source's contribution to the Indonesian NER task. Taking advantage of the cross-lingual

*https://github.com/khairunnisaor/idner-news-2k

transfer method to increase the dataset size of the gold NER Indonesian dataset may also help the model to enhance its performance. Moreover, pre-training an entity-aware language model to increase the robustness of the model when recognizing entities for the Indonesian language would be interesting to explore.

Acknowledgements

I would like to express my gratitude to my supervisor, Assoc. Prof. Mamoru Komachi, for giving me the opportunity to do the Master's study in his laboratory and providing invaluable guidance throughout this research. His patient support and overall insights have carried me through all the stage of doing my study and had me experiencing international collaborative research and conference. I also would like to thank Prof. Takama Yasufumi, Prof. Toru Yamaguchi, Asst. Prof. Eri Shimokawara, and Asst. Prof. Hiroki Shibata for the co-supervision and advice during the final semester evaluations.

I would like to give special thanks to Aizhan Imankulova for giving both academic and mental support as my mentor, as well as the deep-talk discussions that gives me insights to many things. Thank you to Mana Ashida, Naomi Shiraishi, Tosho Hirasawa, Chen Zhousi, and all of the lab members for helping me adapt during my early life in Japan and conducting research in our lab.

Finally, I would like to thank my parents and little brother for the endless support and letting me pursue a higher study far away from home.

References

[1] W. U. Ahmad, Z. Zhang, X. Ma, K.-W. Chang, and N. Peng. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 372–382. Association for Computational Linguistics, 2019.

[2] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[3] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[4] I. Alfina, R. Manurung, and M. I. Fanany. 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, 2016.

[5] R. Artstein and M. Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[6] B. Aryoyudanta, T. B. Adji, and I. Hidayah. 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, 2016.

[7] D. Benikova, C. Biemann, and M. Reznicek. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth In-*

ternational Conference on Language Resources and Evaluation (LREC'14), pages 2524–2531. European Language Resources Association (ELRA), 2014.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[9] I. Budi and S. Bressan. Application of association rules mining to named entity recognition and co-reference resolution for the Indonesian language. *International Journal of Business Intelligence and Data Mining*, 2(4):426–446, 2007.

[10] I. Budi, S. Bressan, G. Wahyudi, Z. A. Hasibuan, and B. A. A. Nazief. Named entity recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach. In A. Hoffmann, H. Motoda, and T. Scheffer, editors, *Discovery Science*, pages 57–69, 2005.

[11] M. Cettolo, N. Jan, S. Sebastian, L. Bentivogli, R. Cattoni, and M. Federico. The IWSLT 2016 evaluation campaign. International Workshop on Spoken Language Translation, 2016.

[12] A. Chaudhary, J. Xie, Z. Sheikh, G. Neubig, and J. Carbonell. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. 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 5164–5174, 2019.

[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[14] A. Conneau and G. Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and

R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 1–11. Curran Associates, Inc., 2019.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[16] L. Duong, T. Cohn, S. Bird, and P. Cook. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122. Association for Computational Linguistics, 2015.

[17] R. Eskander, S. Muresan, and M. Collins. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831. Association for Computational Linguistics, 2020.

[18] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382, 1971.

[19] Y. Fu, N. Lin, X. Lin, and S. Jiang. Towards corpus and model: Hierarchical structured-attention-based features for Indonesian named entity recognition. In *Proceedings of the Journal of Intelligent Fuzzy Systems, vol. Pre-press, no. Pre-press, pp. 1-12, 2021*, 2021.

[20] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487, 2018.

[21] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[22] W. Gunawan, D. Suhartono, F. Purnomo, and A. Ongko. 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, 2018.

[23] J. He, Z. Zhang, T. Berg-Kirkpatrick, and G. Neubig. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3223. Association for Computational Linguistics, 2019.

[24] Z. He, Z. Wang, W. Wei, S. Feng, X. Mao, and S. Jiang. A survey on recent advances in sequence labeling from deep learning models. *CoRR*, abs/2011.06727, 2020.

[25] D. Hoesen and A. Purwarianti. 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, 2018.

[26] Z. Huang, W. Xu, and K. Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv e-prints*, page arXiv:1508.01991, 2015.

[27] F. Ikhwantri. Cross-lingual transfer for distantly supervised and low-resources Indonesian NER. *CoRR*, abs/1907.11158, 2019.

[28] A. Jain, B. Paranjape, and Z. C. Lipton. Entity projection via machine translation for cross-lingual NER. 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 1083–1092. Association for Computational Linguistics, 2019.

[29] B. Ji, Z. Zhang, X. Duan, M. Zhang, B. Chen, and W. Luo. Cross-lingual pre-training based transfer for zero-shot neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):115–122, 2020.

[30] Y. Kim, Y. Gao, and H. Ney. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy, 2019. Association for Computational Linguistics.

[31] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin. 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. International Committee on Computational Linguistics, 2020.

[32] T. Kudo. 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, 2018.

[33] K. Kurniawan, L. Frermann, P. Schulz, and T. Cohn. PPT: Parsimonious parser transfer for unsupervised cross-lingual adaptation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2907–2918. Association for Computational Linguistics, 2021.

[34] K. Kurniawan and S. Louvan. 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, 2018.

[35] K. Kuspriyanto, O. S. Santoso, D. H. Widyantoro, H. Sastramihardja, K. Muludi, and S. Maimunah. Performance evaluation of SVM-based information extraction using margin values. *International Journal on Electrical Engineering and Informatics*, 2:256–265, 2010.

[36] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, page 282–289. Morgan Kaufmann Publishers Inc., 2001.

[37] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 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, 2016.

[38] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, pages 1–14, 2018.

[39] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, pages 1–14, 2018.

[40] R. A. Leonandya, B. Distiawan, and N. H. Praptono. 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, 2015.

[41] R. A. Leonandya, B. Distiawan, and N. H. Praptono. A semi-supervised algorithm for Indonesian named entity recognition. In *2015 3rd International Symposium on Computational and Business Intelligence (ISCBI)*, pages 45–50, 2015.

[42] R. A. Leonandya and F. Ikhwantri. Pretrained 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, 2019.

[43] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2020.

[44] I. Marinova. 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, 2019.

[45] S. Mayhew, C.-T. Tsai, and D. Roth. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545. Association for Computational Linguistics, 2017.

[46] F. Muhammad and M. L. Khodra. Event information extraction from Indonesian tweets using conditional random field. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6, 2015.

[47] J. Ni, G. Dinu, and R. Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480. Association for Computational Linguistics, 2017.

[48] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53. Association for Computational Linguistics, 2019.

[49] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 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, 2018.

[50] T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. Association for Computational Linguistics, 2019.

[51] B. Plank and Ž. Agić. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620. Association for Computational Linguistics, 2018.

[52] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics, 2020.

[53] A. Rahimi, Y. Li, and T. Cohn. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164. Association for Computational Linguistics, 2019.

[54] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora.*, pages 82–94, 1995.

[55] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. M. Soe, K. T. Nwet, M. Utiyama, and C. Ding. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6, 2016.

[56] S. Schuster, S. Gupta, R. Shah, and M. Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805. Association for Computational Linguistics, 2019.

[57] J. Sneddon. *The Indonesian Language: Its History and Role in Modern Society.* University of New South Wales, 2003.

[58] L. Sun, H. Yi, and H. Chen. Back attention knowledge transfer for low-resource named entity recognition. *CoRR*, abs/1906.01183, 2019.

[59] Y. Syaifudin. Quotations identification from Indonsian online news using rule-based method. Master's thesis, Universitas Gadjah Mada, 2016. Undergraduate Thesis.

[60] N. Taufik, A. F. Wicaksono, and M. Adriani. Named entity recognition on Indonesian microblog messages. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*, pages 358–361, 2016.

[61] W. L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953.

[62] J. Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA), 2012.

[63] E. F. Tjong Kim Sang. 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, 2000.

[64] E. F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 1–4, 2002.

[65] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[66] C.-T. Tsai, S. Mayhew, and D. Roth. Cross-lingual named entity recognition via Wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228. Association for Computational Linguistics, 2016.

[67] R. van der Goot, I. Sharaf, A. Imankulova, A. Üstün, M. Stepanović, A. Ramponi, S. O. Khairunnisa, M. Komachi, and B. Plank. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2479–2497. Association for Computational Linguistics, 2021.

[68] X. Wang, Y. Jiang, N. Bach, T. Wang, F. Huang, and K. Tu. Automated concatenation of embeddings for structured prediction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

[69] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857. Association for Computational Linguistics, 2020.

[70] D. C. Wintaka, M. A. Bijaksana, and I. Asror. 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, 2019.

[71] Q. Wu, Z. Lin, B. Karlsson, J.-G. Lou, and B. Huang. 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, 2020.

[72] Q. Wu, Z. Lin, G. Wang, H. Chen, B. F. Karlsson, B. Huang, and C.-Y. Lin. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9274–9281, 2020.

[73] S. Wu and M. Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844. Association for Computational Linguistics, 2019.

[74] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv e-prints*, page arXiv:1609.08144, 2016.

[75] J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, 2018.

[76] V. Yadav and S. Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158. Association for Computational Linguistics, 2018.

[77] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454. Association for Computational Linguistics, 2020.

[78] I. Yamashita, S. Katsumata, M. Kaneko, A. Imankulova, and M. Komachi. Cross-lingual transfer learning for grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4704–4715. International Committee on Computational Linguistics, 2020.

[79] R. Östling and J. Tiedemann. Efficient word alignment with Markov chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*, 106(1):125–146, 2016.

Publication List

International Conferences

[1] Siti Oryza Khairunnisa, Aizhan Imankulova, and Mamoru Komachi. Towards a Standardized Dataset on Indonesian Named Entity Recognition. In the 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: Student Research Workshop. Online. December 4-7, 2020.

[2] Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi and Barbara Plank. 2021. From Masked Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-shot Spoken Language Understanding. In the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online. June 6-11, 2021.