EBERT

By Meilany Nonsi Tentua

ICIC Express Letters ICIC International ©2020 ISSN 1881-803X
Volume **14**, Number **5**, May **2020** pp. 1–8

ENHANCED BIDIRECTIONAL ENCODER TRANSFORMERS WITH RELATIVE POSITION FOR INDONESIAN SKILL RECOGNITION

Meilany Nonsi Tentua [1,3], Suprapto[2,*] and Afiahayati[2]

[1]Doctoral Program, Department of Computer Science and Electronics
[2]Department of Computer Science and Electronics
Faculty of Mathematics and Computer Sciences
Universitas Gadjah Mada,
North Sekip, Bulak Sumur, Yogyakarta, Indonesia
*Corresponding author: sprapto@ugm.ac.id

[3]Departement Informatics
Faculty of Science and Technology
Universitas PGRI Yogyakarta
JL.IKIP PGRI No. 112, Indonesia

Received xxxx; accepted xxxx

Abstract. *This paper presents an approach to improve Indonesian skill recognition using enhanced Bidirectional Encoder Transformers with relative position embeddings (EBERT-RP). The proposed method aims to overcome the challenges in recognizing Indonesian skills due to the complexity of the Indonesian language and the lack of annotated data. The EBERT-RP model incorporates relative position embeddings, which allow the model to capture the relative positions of tokens in a sentence, and a novel attention mechanism that improves the model's ability to attend to important information. To evaluate the performance of the EBERT-RP model, we conducted experiments on a dataset of Indonesian skill recognition task. Our results show that the EBERT-RP model outperforms other state-of-the-art models, achieving an accuracy of 90.1% on the test set. Furthermore, we conducted an ablation study to analyze the contribution of the relative position embeddings and the attention mechanism to the performance of the model. The results show that both the relative position embeddings and the attention mechanism are crucial for achieving high performance.*
Keywords: EBERT-RP, Skill Recognition, Relative Position

1. **Introduction.** Natural Language Processing (NLP) has become increasingly important in recent years, with numerous applications such as text classification [1], sentiment analysis [2], text generation [3], and skill recognition [4]. Skill recognition is a challenging task that involves identifying skills and competencies from a candidate's resume or job application[5]. This task has gained significant attention due to its importance in the recruitment process for many industries [6].

The success of NLP models heavily relies on the ability to capture the relationship between words in a sentence. Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained NLP model that has shown remarkable performance in various tasks [7]. However, BERT does not consider the relative position between words in a sentence [8], which is particularly important for languages with complex grammar and word order such as Indonesian.

The recognition of skills from job postings and resumes is a crucial task in talent management and recruitment processes [9]. Skill recognition is challenging due to the variation in language use, context, and expression of skills across different job postings and resumes [4]. Named Entity Recognition (NER) has been widely used to identify and extract named entities, such as persons, organizations, and locations, from natural language texts. However, conventional NER methods do not perform well in recognizing skills, which often involve complex expressions and domain-specific terminology.

In recent years, deep learning approaches, particularly Bidirectional Encoder Transformers (BERT), have shown promising results in NER tasks [10]. However, existing BERT-based models have limited performance in recognizing skills due to the model's inability to capture the complex relationships and dependencies between tokens in a sequence. The utilization of domain-adaptive pre-training such language is advantageous in enhancing the performance of the task for both hard skills and soft skill components. BERT multilingual [7], IndoBERT [11] and IndoNLU [12] are pretrained model Indonesian language.

There has been a growing interest in incorporating relative position mechanisms into deep learning models for NLP tasks. Several studies have proposed methods to enhance BERT with relative position mechanisms for various NLP tasks, such as NER and relation extraction [13]. These methods have shown improved performance compared to the original BERT model [14]. However, to the best of our knowledge, there has been no study that investigates the use of relative position mechanisms for skill recognition tasks, particularly in the Indonesian language.

In this paper, we propose an enhanced bidirectional encoder transformer with relative position (EBERT-RP) for Indonesian skill recognition. Our model incorporates relative positional encoding to capture the relationship between words in a sentence. We also introduce a new skill recognition dataset for the Indonesian language.

The main contributions of this work are as follows: (1) we propose a new model for Indonesian skill recognition using enhanced bidirectional encoder transformers with relative position. The enhancement involves incorporating relative position information into the model architecture. Relative position information refers to the position of words relative to each other in a sentence or sequence of words, rather than absolute positions in the input sequence; (2) we evaluate our proposed model on the benchmark dataset and compare it with other baseline models, demonstrating its superior performance.

2. **Literature Review.** This section describes about the research conducted in skill recognition and the relative position embedding in transformers.

2.1. Relative position embedding in Transformers. Transformers [15] that use self-attention mechanisms have been adopted in various NLP tasks due to their parallelism and excellence in modeling very long contexts. Relative position embedding is a Transformer positional information proposed to improve the weaknesses of absolute position embedding with sinusoidal function. Relative position was first introduced by adding a vector as directional information from the input element [16]. This vector is embedded in the key matrix in calculating attention values and the value matrix in calculating attention filtering values. It was further developed in Transformer-XL [17] and XLNet [3]. The relative position embedding was improved in Transformer-XL and XLNet by adding a bias parameter in the form of a vector to the content-based and location-based attention. The relative position embedding proposed in Transformer-XL and XLNet was used in NER modeling [13]. The modification made was to remove the matrix parameter and maintain the bias parameter.

2.2. Skill Recognition. Several studies have addressed skill recognition tasks using various methods. BERT-BiLSTM-CRF has been used in studies for skill recognition tasks [18] in English language. It was found that, the model based on BERT pretraining vector was better. However, the disambiguation of multi-sense skills is recognition and normalization of occupational skills in online recruitment. Using word embedding to quantify skills apply Markov Chain Monte Carlo (MCMC) methods to aggregate vectors into clusters that represent respective senses [19]. That clustering algorithm shows outperforms from other clustering algorithms for the disambiguation problem.

The automated approach for skill entity recognition and normalization has important applications in workforce training and job matching, where it can significantly improve the accuracy of identifying and matching relevant skills with skill taxonomi generation and skill tagging [6]. The problem being addressed is an extreme multi-label classification (XMLC) problem[9] and SVM [20], where the binary relevance of thousands of individual skills needs to be determined based on the descriptions provided. The model effectively tackles the issue of missing skills and can help recover relevant skills that may have been overlooked during the job posting process.

SkillNER enables the detection of communities of job profiles based on their shared soft skills and communities of soft skills based on their shared job profiles [4]. This system demonstrates that can automatically retrieve soft skills from a large corpus in an efficient way, proving useful for firms, institutions, and workers.

Deep learning methods, such as BERT, can be effective for skill recognition tasks in various languages. Using domain-adaptive pre-training is beneficial in improving the performance of the task in terms of both hard skills and soft skill components. Several study for skill recognition use domain language adaptive pre-trained susch as, Finnish language (FinBERT) [21], and JobSpanBERT [22].

3. **Material and Methodology.** In this section, we present our approach for Enhanced Bidirectional Encoder Transformer with Relative Position (EBERT-RP) for Indonesian skill recognition, which is depicted in Figure 1. Our model comprises of two modules, namely pretrain model language EBERT-RP and skill recognition modelling. Each module have five step: tokenization, preprocessing, propose model of EBERT-RP and training model.

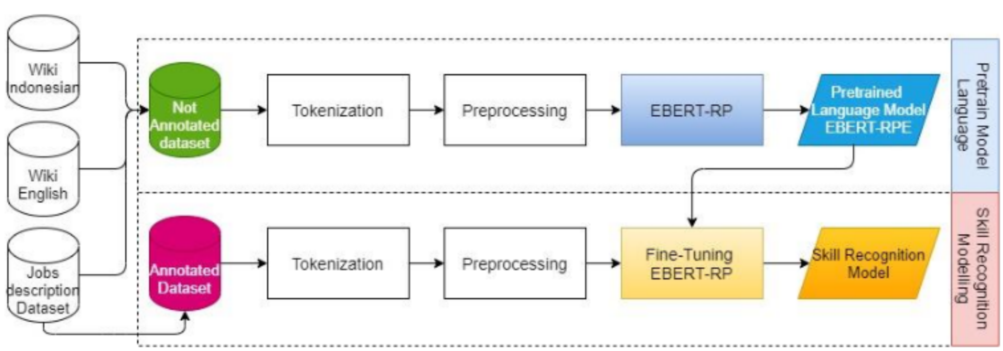

FIGURE 1. Architecture of EBERT-RP for skill recognition.

3.1. **Material**. The EBERT-ERP pre-training process and skill recognition modeling used a dataset composed of Indonesian language job requirements gathered from various job portals. The collected data was then filtered to remove duplicates, resulting in 34,966 job requirements. The dataset was then divided into two groups: one for pre-training the language model and the other for Skill recognition modelling.

The dataset used for pre-training the EBERT-RP language model did not include any annotations. To enhance the corpus for pre-training EBERT-RP we add the Indonesian Wikipedia corpus and the English Wikipedia corpus to job requiretments dataset.

The dataset used in the Skill recognition modelling follows the BIO tagging format [23] which includes three labels: Beginning, Inside, and Other. The dataset contains 4,394 rows of job requirements that have been pre-processed through word tokenization and manually labeled by an annotator to identify entities related to B-HSkill, S-Skill, B-Tech, I-HSkill, I-SSkill and I-Tech.

3.2. **Tokenization**. Tokenize the text data into word pieces, also known as subwords. This is done using an algorithm called WordPiece, which breaks words into smaller units based on the frequency of occurrence. The vocabulary set in the EBERT-ERP tokenizer is a set of all the unique tokens that the model uses to represent text data. These tokens are the building blocks of the input sequences that are fed into the EBERT-ERP model. The vocabulary set includes a special set of tokens, such as the [CLS] token to indicate the start of a sequence, the [SEP] token to separate sentences or different parts of a sequence, and the [MASK] token to mask out certain tokens during training. The size of the vocabulary set in EBERT-ERP model has 31,923 of tokens. Table 1 shows sample of tokenizing a job description.

TABLE 1. Sample of tokenizing a job description

Job description	Tokenize
Mahir dalam bahasa SQL, HTML, dan VB.	['[CLS]', 'mahir', 'dalam', 'bahasa', 'sql', ',', 'html', ',' , 'dan', 'vb', '.', '[SEP]']

3.2. **Preprocessing**. The job requirements dataset is preprocessed by tokenizing the text into individual words using the WordPiece tokenizer. The resulting tokens are then converted to their corresponding token IDs and segment IDs, as well as masked to indicate which tokens are actual words and which are padding. We also add special tokens to indicate the start and end of each sentence in the text. Table 2 show sample of preprocessing a job description.

TABLE 2. Sample of preprocessing a job description

Sentence	Mahir dalam bahasa SQL, HTML, dan VB.
Token IDs:	[2, 24633, 1878, 2760, 12271, 16, 10727, 16, 1622, 30020, 18, 3]
Segment IDs:	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Mask IDs:	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

The integer IDs assigned to each word or subword token in the input text will be referred to as the input word vector t_i and will be added to the absolute position vector p_i. Maximum words can be process in EBERT-RP is $n = 256$. We use Equation (1) and (2) for absolute position vector p_i.

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d}) \tag{1}$$

$$PE_{(pos,2i+1)} = \sin(pos/10000^{2i/d}) \tag{2}$$

The result of the addition will produce vector x_i. Equation (3) shows the addition process on each word vector.

$$x_i = t_i + p_i \tag{3}$$

Where x_i= output vector in the position embedding layer ; t_i= word vector ; p_i= absolute position vector

3.4. **Model architecture**. The EBERT-RP model is based on the BERT architecture but includes several modifications. We add a relative position embedding layer to the model to help it recognize the relative positions of skills in the text. They also modify the input layer to include a token position embedding layer, which helps the model recognize the positions of individual words in the text.

A sequence input $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ with $x_i \in \mathbb{R}^{d_x}$ and m is the length of the sentence will be projected using three matrices $W^Q \epsilon\ \mathbb{R}^{mxd_q}$, $W^K \epsilon\ \mathbb{R}^{mxd_k}$, and $W^V \epsilon\ \mathbb{R}^{mxd_v}$ to extract feature representations Q, K, and V, which are called query, key, and value with $d_k = d_q$. The Q, K, and V matrices are calculated using equation (4).

$$Q = x_i W^Q \ , \ K = x_i W^K \quad , \ V = x_i W^V \tag{4}$$

Relative positional embedding is done by adding a vector $a_{ij}^V, a_{ij}^K \in \mathbb{R}^{d_a}$, where a_{ij}^V and a_{ij}^K are the direction information of the input elements x_i and x_j. The vector a_{ij}^K is embedded in the matrix K = $x_j W^K$ so that it will interact with matrix Q = $x_i W^Q$ in calculating the attention score. Meanwhile, vector is embedded in matrix V = $x_j W^V$ in calculating the attention filtration. The modified equations (5) and (6) with relative positional embedding. Equation (7) shows the interaction of relative position with matrix Q.

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + a_{ij}^V) \tag{5}$$

$$e_{ij} = \frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}} \tag{6}$$

$$e_{ij} = \frac{x_i W^Q (x_j W^K)^T + x_i W^Q a_{ij}^{K^T}}{\sqrt{d_z}} \tag{7}$$

3.4. **Training**. We train EBERT-RP models from scratch based on the aforementioned configuration. We train EBERT-RP based on a MLM objective as previous research has done [7], randomly we select 15% of tokens, and then substitute 80% of these tokens with "[MASK]", subtitute another random 10% token , and take care 10% of the tokens unspoiled. We use transformers encoder with 12 hidden layers (dimension = 768), 12 attention heads, and 3 hidden feedforward layers (dimension = 3,072). The only difference is the maximum sequence length, fixed at 256 tokens based on the average number of tokens in job requests.

3.5. **Fine-tuning**. The EBERT-RP pretrained language model is used to identify skill entities in job requirements. A linear layer and softmax activation are added in the NER Skill model to identify entities. We use one of the hyperparameters suggested in Devlin et al.'s research (2019) as the optimal hypermeter value in fine-tuning the model, such as learning rate is 5e-3, epoch = 1-5 and batch size = 32.

4. **Result and Discussion.** The EBERT-RP language models were trained using the NVIDIA A100-SXM machine. The training process used hyperparameters such as a learning rate of 1e-04, a dropout rate of 0.1, a batch size of 64, and 300 epochs with 448.500 steps. The resulting model have 110M parameters.
The EBERT-RP model is then fine-tuned on the annotated job requirements dataset for modelling Indonesian skill recognition. We use a batch size of 32 and train the model for 11 epochs with an Adam optimizer.
We evaluate the performance of our proposed model on the test set using precision, recall, F1-score at the level of individual tokens and entities. Table 3 shows performance model at the token level. The token "O" has the highest performance (F1 score =96%) , as it is owned by most of the tokens in the dataset because this entity. Token "B-Tech" have F1 score= 85%, because this entity is commonly owned by a single word. Overall this model have performance F1-Score= 74% at token level.

TABLE 3. Performance model at the token level

TOKEN	precision	recall	f1-score
B-HSkill	0.69	0.53	0.60
B-SSkill	0.76	0.86	0.81
B-Tech	0.88	0.81	0.85
I-HSkill	0.72	0.51	0.60
I-SSkill	0.80	0.54	0.64
I-Tech	0.83	0.60	0.69
O	0.94	0.97	0.96

Table 4 shows performance model at the entity level. The token "B-SSkill" has the highest performance (F1 score =97%) , and the lowest performance is "I-Tech" (F1 score = 82%). The proposed model achieved an F1-score of 90.1% at entity levels.

TABLE 4. Performance model at the entity level

ENTITAS	precision	recall	f1-score
B-HSkill	0.88	0.92	0.90
B-SSkill	0.95	0.98	0.97
B-Tech	0.94	0.95	0.94
I-HSkill	0.89	0.87	0.88
I-SSkill	0.93	0.86	0.90
I-Tech	0.88	0.76	0.82

We compare the performance of our model with other baseline pretrained language models, including BERT without relative position embedding, BERT [7], IndoBERT [11], IndoNLU [12]. Table 5 shows performace comparation between pretrained model Indonesian language.

TABLE 5. Comparation performance model

		Token Level			Entity Level		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Baseline	BERT	81%	75%	78%	87.9%	85.9%	86.3%
	IndoNLU	52%	30%	31%	58.3%	49.0%	53.2%
	IndoBERT	81%	76%	78%	87.4%	85.7%	86.5%
Ours	EBERT-RP	80.2%	68.8%	73.5%	**91.1%**	**89.0%**	**90.1%**

The proposed model achieved an F1-score of 73.5% for token level and 90.1% for entity level, outperforming the baseline models. The results showed that adding relative position embedding can significantly improve the performance of the model in recognizing Indonesian skills.

5. Conclusion. In this paper, we propose Enhanced Bidirectional Encoder Transformers with Relative Position (EBERT-RP) for Indonesian skill recognition. We incorporate a relative positional encoding layer into the BERT architecture to enhance the model's ability to capture the relationship between words in a sentence.

Our experimental results demonstrate that EBERT-RP outperforms baseline models such as BERT [7], IndoNLU [12] and IndoBERT [11] in terms of precision (91.1%), recall(89%) and F1-score (90.1%) for entity level. The ablation study shows that the relative positional encoding layer contributes significantly to the model's performance.

Our work contributes to the development of natural language processing techniques for low-resource languages such as Indonesian. The incorporation of relative positional encoding is a promising approach for improving the performance of language models in languages with complex syntax and grammar.

EBERT

ORIGINALITY REPORT

10%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.researchgate.net Internet	106 words — 4%
2	www.icicel.org Internet	30 words — 1%
3	www.coursehero.com Internet	18 words — 1%
4	Zahra Maria, Allison R. Campolo, Benjamin J. Scherlag, Jerry W. Ritchey, Véronique A. Lacombe. "Dysregulation of insulin-sensitive glucose transporters during insulin resistance-induced atrial fibrillation", Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2018 Crossref	16 words — 1%
5	Bijian Jian, Chunbo Ma, Dejian Zhu, Yixiao Sun, Jun Ao. "Seeing through Wavy Water–Air Interface: A Restoration Model for Instantaneous Images Distorted by Surface Waves", Future Internet, 2022 Crossref	15 words — 1%
6	bmcmedinformdecismak.biomedcentral.com Internet	12 words — < 1%
7	www.mdpi.com Internet	12 words — < 1%

8 Haofeng Li, Junjia Huang, Guanbin Li, Zhou Liu, Yihong Zhong, Yingying Chen, Yunfei Wang, Xiang Wan. "View-Disentangled Transformer for Brain Lesion Detection", 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022 — 11 words — < 1%
Crossref

9 Yutong Wu, Han Qiu, Tianwei Zhang, Jiwei Li, Meikang Qiu. "Watermarking Pre-trained Encoders in Contrastive Learning", 2022 4th International Conference on Data Intelligence and Security (ICDIS), 2022 — 11 words — < 1%
Crossref

10 earsiv.etu.edu.tr — 11 words — < 1%
Internet

11 "Extensive and Intensive: A BERT-based machine reading comprehension model with two reading strategies", Academic Journal of Computing & Information Science, 2021 — 10 words — < 1%
Crossref

12 ipfs.io — 9 words — < 1%
Internet

13 Haodong Wang, Jun Xie, Xinying Xu, Zihao Zheng. "Few-Shot PCB Surface Defect Detection Based on Feature Enhancement and Multi-Scale Fusion", IEEE Access, 2022 — 8 words — < 1%
Crossref

14 paleo.sscnet.ucla.edu — 8 words — < 1%
Internet

15 www.ncbi.nlm.nih.gov — 7 words — < 1%
Internet

16 Qinlong Luo, Meng Zhao, Faizan Javed, Ferosh Jacob. "Macau: Large-scale skill sense disambiguation in the online recruitment domain", 2015 IEEE International Conference on Big Data (Big Data), 2015 — 6 words — < 1%
Crossref

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF