Multilingual Named Entity Recognition Model for Indonesian Health Insurance Question Answering System

Budi Sulistiyo Jati
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
budi.sulistiyo.j@mail.ugm.ac.id

Widyawan, ST, M.Sc., Ph.D.
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
widyawan@ugm.ac.id

Muhammad Nur Rizal, S.T., M.Eng., PhD
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
mnrizal@ugm.ac.id

2020 3rd International Conference on Information and Communications Technology (ICOIACT) | 978-1-7281-7356-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICOIACT50329.2020.9332027

***Abstract*—Named Entity Recognition (NER) is the task of extracting information to find and classify entities from unstructured text into predetermined categories. In this study, NER is used to find entities of locations, organizations, financial tasks, administrative tasks, and healthcare facilities in chat and public service complaints dataset of Indonesian national health insurance. The method used is Bidirectional Encoder Representation from Transformer (BERT) Multilingual Cased, and BERT Multilingual Uncased models. Pre-processing conducted in this research is tokenization, formalization, and tag distribution analysis. Then it is converted into a BERT input feature consisting of token embedding, positional embedding, and attention mask. Based on the experiment results, BERT Multilingual Uncased model archives total average F1 score 83.52 and BERT Multilingual Cased model archives total average F1 score 85.41. The experiment results prove that BERT Multilingual can be implemented for Indonesian dataset, and also show that the cased model can get a better F1 score.**

Keywords—named entity recognition, natural language processing, information extraction, BERT

I. INTRODUCTION

National health insurance services must be easily accessible when people need information, have problems, or want to provide feedback about existing health services. One way to improve public service is to use customer service, but customer service has operational time limits so that service information can be delayed. A chatbot is an artificial intelligence technology that is embedded in automated instant messages that are often used by large companies as virtual agents that can communicate with customers and serve customers as human beings. With an automated chatbot, services can run at any time without any operational time limits.

Chatbot has many benefits for the government because it can provide better services and information and save labor costs by providing services 24/7. Chatbot that is used in public services also encourages public participation by gathering feedback from citizens. This feedback is then used for the purpose of decision making as the information needs of citizens [1]. Large companies also use chatbot as a service facility because it can increase profits, efficiency, functionality, and sustainability. Chatbot can personally reach customers using natural language with the support of artificial intelligence. The response generated by the chatbot is the result of a keyword scan on the input made by the user. Input produces a response that is considered the best reply, so the conversation seems to be carried out by two people communicating with each other.

To find keywords in an unstructured document, named entity recognition can be used. Many Natural Language Processing (NLP) applications require entity discovery in textual documents. Entity names can be people, companies, locations, times, etc. The task of identifying entities in text is called Named Entity Recognition (NER) and is often used in information extraction. Information extraction is the process of finding information of a document or natural language that produces useful information in the form of structured information with a certain format. The main task of NER is to identify and classify entities in unstructured datasets into predetermined classes.

Some approaches to detect named entities are the classic approaches using rule-based, modern approaches using machine learning, and using deep learning. To develop deep learning methods with high performance, the choice of neural network type needs to be considered because the text is a sequential data format. LSTM (Long Short Term Memory) is a network that can process sequential input and update a kind of vector state that contains information about all the past elements that can make predictions from that information. For NER, the context includes sequential past and future labels, so it is necessary to consider past and future information. Bidirectional LSTM is a combination of two LSTMs, one running forward from right to left and one running backwards from left to right.

A recent paper published by Google AI-Language researchers found BERT (Bidirectional Encoder Representation from Transformer) that provides state-of-the-art results in various NLP tasks, such as Question Answering System or Chatbot and Natural Language Inference [2]. For named entity recognition research in Indonesian language, LSTM network has been widely implemented. However, there is no experiment on named entity recognition of Indonesian language using BERT especially for chat and public service complaints dataset. In this research, we will apply named entity recognition in Indonesian language using BERT in a case study of a chat and public service complaints in Indonesian health care insurance. Experiments will also compare using cased and uncased models to see whether maintaining uppercase letters affects the results.

II. EXISTING WORK

Several studies have reached state-of-the-art presenting a new neural network architecture that automatically detects word-level and character features using Bidirectional LSTM

and CNN architecture. The evaluation results showed a neural network model created achieving state-of-the-art results for named entity recognition. The model was capable of studying complex relationships of large amounts of data with a F1 score 91.62 for CoNLL-2003 data, and F1 score 86.28 for the OntoNotes dataset [3].

The other research is by comparing various LSTM networks to sequence tagging. The models are LSTM, LSTM-CRF, Bidirectional LSTM, and Bidirectional LSTM-CRF. Results showed that the Bidirectional LSTM combined with the CRF could use the past and future inputs efficiently due to LSTM's Bidirectional components. This Model can also use sentence-level tag information because of the CRF layer. The performance of the model archives resulted in F1 score 90.10 for the CoNLL 2003 DataSet [4]. Similar research is also conducted and achieves results in F1 score 90.90 which uses a new form of learning embedding of words that can utilize information from the relevant lexicon to enhance representation [5].

The Bidirectional LSTM-CRF model is also used for Russian language case studies. The study compared three models, that is Bidirectional LSTM, Bidirectional LSTM-CRF, and Bidirectional LSTM-CRF added with external word embedding [6]. The three models were evaluated using Gareev, Person-1000, and FactRuEval-2016 dataset. The results show the Bidirectional LSTM-CRF model significantly improves the quality of predictions.

Since the publication of a paper about BERT with state-of-the-art results in various fields of NLP [2], there have been many studies on named entity recognition using BERT in various fields. One of the research conducted by evaluated the BERT Multilingual model (mBERT) in German and English language [7]. Research extends previous work on fine-tuning language models by applying them to the BERT architecture [8]. The results show that the generalized multilingual model works well for NER in the selected languages both German and English.

BERT is also used in the field of biomedicine for clinical documents. The study was conducted by evaluating two baselines based on BERT Multilingual and BioBERT on the PharmaCoNER corpus to recognize chemical and protein entities from the Spanish biomedical text [9]. The BERT Multilingual model has F1 score 89.24, while the BioBERT model produces F1 score 89.02. This is because BioBERT is only pre-trained for English biomedical texts. It is considered that a large number of chemicals and proteins have the same name in English and Spanish in the biomedical literature. While BERT Multilingual uses a large-scale dataset in which there are 104 languages with various domains in pre-training. The experimental results show that transferring knowledge learned from large-scale dataset sources on BERT Multilingual to the target domain provides an effective solution for PharmaCoNER tasks.

Other studies comparing four pre-training models, namely BERT, ERNIE, ERNIE2.0-tiny, and RoBERTa [10]. ERNIE is a pre-training model that uses entity levels and level phrases in masking strategies to get language representations [11]. While ERNIE2.0-tiny is the result of compressing from ERNIE 2.0, a continual pre-training framework that can gradually build and train various pre-training tasks through continuous multi-task learning [12]. RoBERTa is similar to BERT, by changing the masking strategy from static to dynamic and removing NSP (Next Sentence Prediction) task [13]. Comparison results in the MSRA-2006 dataset indicate that RoBERTa produced the highest F1 scores with each score for BERT 93.30, ERNIE 93.37, ERNIE2.0-tiny 86.52, RoBERTa 94.17.

NER research for Indonesian datasets has been done by many researchers. One of them is comparing Bidirectional LSTM with Bidirectional LSTM-CNN to look for four different classes, People, Organization, Location, and Event [14]. The results showed the Bidirectional LSTM-CNN network had the highest F1 score of 79.43. The score is considered quite low due to the small number of datasets. Other research uses Bidirectional LSTM-CRF to recognize the entity Person, Location, Organization in the Twitter dataset [15]. The corpus tested consisted of 350 informal tweets and 250 formal tweets with a total of 600 Indonesian tweets. The main challenge in classifying entities in the Twitter dataset is the limited number of words in tweets and the use of informal and uncontrolled grammar, therefore CRF is used for the classification process. The model built gets the best F1 score by adding the word embedding type called FastText with an F1 score 86.13 for formal tweets, F1 score 81.17 for informal tweets, and F1 score 84.11 for combined tweets.

Bidirectional LSTM-CRF is also used in research with a dataset of Indonesian news articles [16]. Bidirectional LSTM is used to combine the previous context and the context afterward by processing data from two directions which are then classified using CRF. Based on the test results using training data of 25,709 words and testing 9,406, the Bidirectional LSTM-CRF method obtained an accuracy of 87.77%.

From previous research that has been summarized, Bidirectional LSTM is widely used and has high F1 score both for the English and Indonesian datasets. Whereas BERT which is rated as getting state-of-the-art in various NLP fields has never been evaluated for Indonesian datasets. For this reason, this research will conduct a study on BERT for named entity recognition in the Indonesian dataset, specifically for the chat and public complaints dataset of Indonesian health care insurance.

III. Experiment Method

The overall structure of the NER experiment is illustrated in Fig. 1. We followed the same structure in the BERT paper for testing German and English [7], but differed in the data pre-processing process. Data will be converted into BERT input features consisting of token embedding, positional embedding, and attention mask. Then the model will be evaluated by measuring the results of the confusion matrix.

A. Datasets

The data source used in this study is a chat dataset on national health insurance combined with a dataset of public complaints from lapor.go.id, a national public service complaint management system created by the Indonesian government as a door to the national complaints channel. The dataset of public service complaints was added because it has the same characteristics and topics as the chat dataset and to be added to the vocabulary so that the total dataset being processed is 3924 row dataset. Information that can be used as a class entity in this study are:

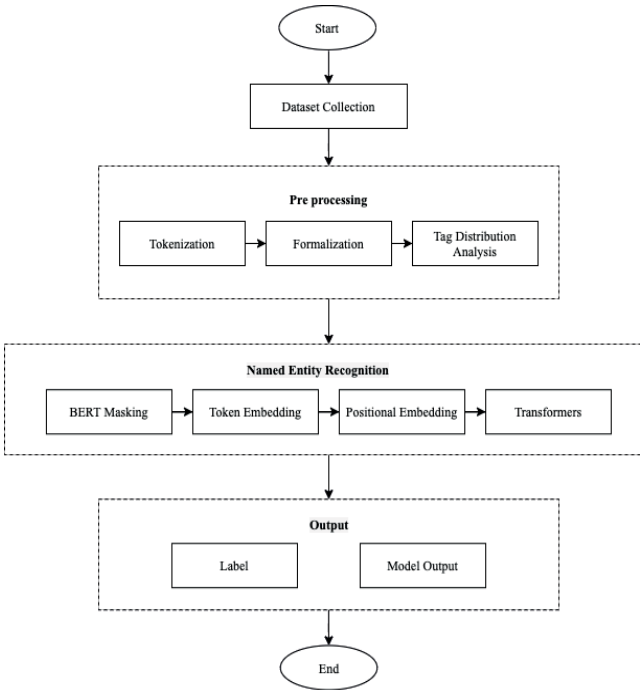

Fig. 2. NER Model Architecture Using BERT

- Organization (ORG): This information includes the name of the organization, including profit-oriented organizations such as companies, factories, etc. and non profit organizations.
- Location (LOC): This information includes names of cities, roads, districts, villages, parks, such as Bandung and Monjali.
- Financial Tasks (FIN): This information includes all matters relating to financial affairs such as bills, fees, etc.
- Administration Tasks (ADMIN): This information includes all matters relating to the administration of services such as registration, classes of healthcare facilities, etc.
- Healthcare Facilities (FASKES): This information contains the types of existing health facilities such as public health center, hospitals, pharmacies, etc.

B. *Preprocessing*

Before entering the named entity recognition stage, pre processing needs to be done. Three processes are conducted in the pre processing stage, that is tokenization, formalization, and BIO Notation. The following is an explanation for each component.

- Tokenization: In this stage, data in the form of sentences are divided into a set of tokens using spaces as separating characters. This is done because the entity classification is done on the token.
- Formalization: At this stage, tokens are converted to formal forms in accordance with Indonesian language standards.
- BIO Notation: The BIO scheme (Beginning, Inside, Outside) is a common tagging format for marking sentences for NER. Here the prefix B indicates that the tag is at the beginning of each chunk.

TABLE 1. EXAMPLE OF LABELING

Conversation	Tag	Entity
Saya	PRP	O
sudah	MD	O
membayar	VB	O
iuran	NN	B-FIN
BPJS	NNP	B-ORG
melalui	X	O
bank	NN	B-ORG
bni	NNP	I-ORG
akan	MD	O
tetapi	CC	O
masih	MD	O
ada	JJ	O
tagihan	NN	B-FIN

The prefix I is for inside of each chunk and the prefix O is for words that do not have an entity in the chunk.

In the labeling example above, there is the word "iuran" with a financial entity and after BIO Tagging becomes B-FIN. Another example is a “bank bni” with organization entity, after BIO Tagging become B-ORG I-ORG.

C. *Named Entity Recognition using BERT*

BERT is unsupervised learning designed to study two-way representation in depth by conditioning the left and right side contexts on all layers [2]. BERT uses bidirectional training from Transformer to language models. This new method can explore deeper language contexts. Previously trained BERT can be adjusted to create new models for various NLP tasks such as relation extraction, question answering system, and named entity recognition. BERT architecture can be seen in Fig. 2.

To find named entity recognition using BERT, in this study the first thing to do is to load a dataset where every word in a sentence is marked with a label. Then do the tag distribution analysis. Because in this experiment the NER process will be carried out as multi-class classification, it is necessary to make training data in the form of token labels so that tokens and labels are obtained. Then the data is organized into training embeddings. After the data is ready, two embedding types are applied, namely token embedding by mapping tokens to id, and positional embedding that is embedding at the position of each token.

The next step is to do a training model. We divide the data into 70% for the training and 30% for the testing. In the trained model, the BERT pre-training model is used, and then fine-tuning is performed. BERT provides two types of models, case models and uncased models. Cased models will leave tokens with uppercase letters, while uncased models will change all tokens to lowercase letters.

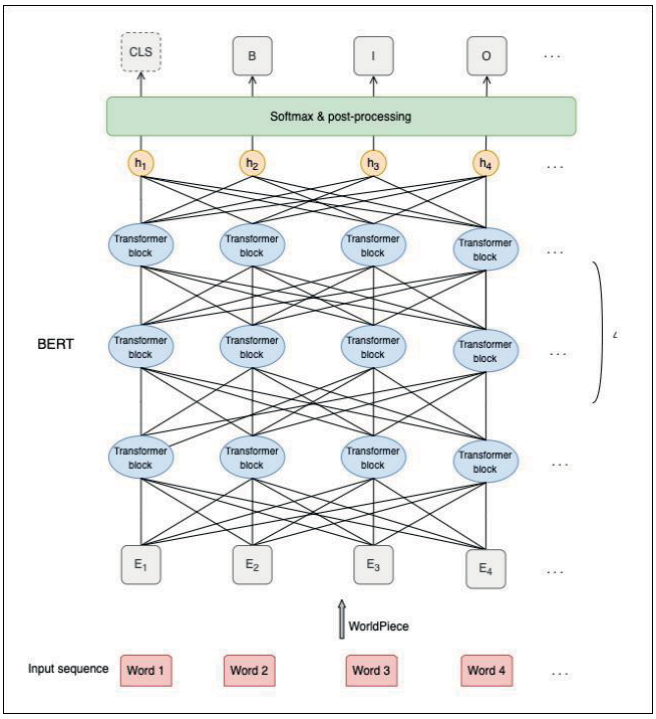

Fig. 2. BERT Architecture (Figure Source: Devlin et al. [2])

Both types of models will be tested to determine the effect of capitalization. Because there is no specific BERT model for Indonesian, we will use BERT Multilingual which includes 104 languages, including Indonesian language. The final step is evaluating the performance of the model. After training a new model for NER, an evaluation of the model needs to be done. Evaluation data can be arranged in the process when organizing previous training data sets. In this study, 30% of the existing datasets are used to test data to determine the performance of the model.

IV. Result and Discussion

The final results of this study obtained the score of the confusion matrix to calculate the performance of classification models such as accuracy, precision, and recall. The evaluation matrices used to measure performance are Precision, Recall, and F1 score.

Table II shows the results of precision, recall, and F1 score for each named entity class in testing using BERT Multilingual Uncased. Table III shows the results of precision, recall, and F1 score for each named entity class in testing using BERT Multilingual Cased.

From the experiments conducted, the total average F1 score obtained from the BERT Multilingual Uncased is 83.52 with an accuracy score of 96.61. Experiments using BERT Multilingual Cased get an total average F1 score 85.41, with an accuracy score of 96.78. This proves that the use of capital letters is important in named entity recognition because in Indonesian language there are homonyms and homographs which are groups of words that have the same writing but have different meanings. So to get the appropriate context of the word, the use of uppercase letters is maintained.

TABLE 2. BERT Multilingual Uncased Result

Named Entity	Precision	Recall	F1 score
LOC	95.35	97.62	96.47
ORG	85.71	95.74	90.45
FIN	95.70	92.86	95.12
FASKES	69.23	45.00	54.55
ADMIN	82.05	80.00	81.01

TABLE 3. Multilingual Cased Result

Named Entity	Precision	Recall	F1-score
LOC	91.11	97.62	94.25
ORG	91.92	96.81	94.30
FIN	95.24	95.24	95.24
FASKES	78.57	55.00	64.71
ADMIN	75.00	82.50	78.57

V. Conclusion

In this paper we introduce the BERT benchmarks for Indonesian, specifically for the chat and public service complaints dataset from Indonesian national health insurance. Research shows that BERT can be used to detect Indonesian entities with results using BERT Multilingual Uncased with total average F1 score of 83.52, and BERT Multilingual Cased with total average F1 score of 85.41. This shows that BERT Multilingual Cased has a better performance in detecting entities and maintaining capital letters in NER is very influential on the results..

For further research, it is necessary to study BERT for Indonesian with more complex dataset such as the Twitter dataset where there is irregular grammar and there is often a mixture of languages in tweets. In addition, it is also necessary to consider building an Indonesian BERT model with a closed domain such as education, health, economy, etc.

References

[1] Y. Petriv, R. Erlenheim, V. Tsap, I. Pappel, and D. Draheim, "Designing Effective Chatbot Solutions for the Public Sector: A Case Study from Ukraine," *Communications in Computer and Information Science*. pp. 320–335, 2020, doi: 10.1007/978-3-030-39296-3_24.

[2] [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv [cs.CL]*. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805.

[3] [3] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *arXiv [cs.CL]*. 2015, [Online]. Available: http://arxiv.org/abs/1511.08308.

[4] [4] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *arXiv [cs.CL]*. 2015, [Online]. Available: http://arxiv.org/abs/1508.01991.

[5] [5] A. Passos, V. Kumar, and A. McCallum, "Lexicon Infused Phrase Embeddings for Named Entity Resolution," *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. 2014, doi: 10.3115/v1/w14-1609.

[6] [6] T. A. Le, The Anh Le, M. Y. Arkhipov, and M. S. Burtsev, "Application of a Hybrid Bi-LSTM-CRF Model to the Task of Russian Named Entity Recognition," *Communications in Computer and Information Science*. pp. 91–103, 2018, doi: 10.1007/978-3-319-71746-3_8.

[7] A. Baumann, Trinity College Dublin, Dublin, and Ireland, "Multilingual Language Models for Named Entity Recognition in German and English," *Proceedings of the Student Research Workshop Associated with RANLP 2019*. 2019, doi: 10.26615/issn.2603-2821.2019_004.

[8] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, doi: 10.18653/v1/p18-1031.

[9] C. Sun and Z. Yang, "Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task," *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 2019, doi: 10.18653/v1/d19-5715.

[10] Y. Wang, Y. Sun, Z. Ma, L. Gao, Y. Xu, and T. Sun, "Application of Pre-training Models in Named Entity Recognition," *arXiv [cs.CL]*. 2020, [Online]. Available: http://arxiv.org/abs/2002.08902.

[11] Y. Sun *et al.*, "ERNIE: Enhanced Representation through Knowledge Integration," *arXiv [cs.CL]*, Apr. 19, 2019.

[12] Y. Sun *et al.*, "ERNIE 2.0: A Continual Pre-training Framework for Language Understanding," *arXiv [cs.CL]*. 2019, [Online]. Available: http://arxiv.org/abs/1907.12412.

[13] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv [cs.CL]*. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692.

[14] W. Gunawan, D. Suhartono, F. Purnomo, and A. Ongko, "Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs," *Procedia Computer Science*, vol. 135. pp. 425–432, 2018, doi: 10.1016/j.procs.2018.08.193.

[15] D. C. Wintaka, M. A. Bijaksana, and I. Asror, "Named-Entity Recognition on Indonesian Tweets using Bidirectional LSTM-CRF," *Procedia Computer Science*, vol. 157. pp. 221–228, 2019, doi: 10.1016/j.procs.2019.08.161.

[16] H. Permana, "Named Entity Recognition Menggunakan Metode Bidirectional Lstm-Crf Pada Teks Bahasa Indonesia," Universitas Komputer Indonesia, 2019.