Received: February 17, 2020. Revised: March 15, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Hybrid Conditional Random Fields and K-Means for Named Entity Recognition on Indonesian News Documents

Joan Santoso[1,3]* Esther Irawati Setiawan[1,3] Eko Mulyanto Yuniarno[1,2]
Mochamad Hariadi[1,2] Mauridhi Hery Purnomo[1,2]*

[1]*Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[2]*Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[3]*Department of Informatics, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia*
* Corresponding author's Email: joan.santoso13@mhs.ee.its.ac.id; hery@ee.its.ac.id

Abstract: The hybrid approach has been widely used in several Natural Language Processing, including Named Entity Recognition (NER). This research proposes a NER system for Indonesian News Documents using Hybrid Conditional Random Fields (CRF) and K-Means. The hybrid approach is to try incorporating word embedding as a cluster from K-Means and take as a feature in CRF. Word embedding is a word representation technique, and it can capture the semantic meaning of the words. The clustering result from K-Means shows that similar meaning word is grouped in the cluster. We believe this feature can improve the performance of the baseline model by adding the semantic relatedness of the word from the cluster features. Word embedding in this research uses Indonesian Word2Vec. The dataset is consisting of 51,241 entities from Indonesian Online News. We conducted some experiments by dividing the corpus into training and testing dataset using percentage splitting. We used 4 scenarios for our experiments, which are 60-40, 70-30, 80-20, and 90-10. The best performance for our model was achieved in 60-40 scenario with F1-Score around 87.18% and also improves about 5.01% compared to the baseline models. We also compare our proposed methods with several models, which are BILSTM and BILSTM-CRF, from previous research. The experiments show that our model can achieve better performance by giving the best improvement of around 4.3%.

Keywords: Named entity recognition, Word2Vec, Hybrid approach, CRF, K-means, Indonesian.

1. Introduction

The Hybrid approach has been implemented in Natural Language Processing Tasks. Hybrid models have been proven to improve the performance of various models. Suncong et. al.[1] uses a hybrid LSTM and CNN to obtain entities and its relation. Another hybrid model in sentiment analysis was combining Gradient Boosting Decision Tree and Support Vector Machine in [2]. A Hybrid method was also developed for Summarization of microblog posts [3] and Sentiment Analysis of Political Data [4]. The main reason why a hybrid approach is chosen as proposed methods is it can combine each model's strength to obtain better performance.

Named Entity Recognition (NER) task has already been developed with Hybrid Model as in [5] and [6]. NER is a task to obtain several types of entities, such as person, location, or organization from text [7]. It has been widely used in several Natural Language Research applications, such as Fake News Detection [8], Aspect Based Sentiment Analysis [9] and Machine Translation [10]. There are various other approaches for obtaining named entity, such as Conditional Random Fields (CRF) [11, 12], Long Short Term Memory (LSTM) [13], moreover with Bidirectional Long Short Term Memory (BILSTM) [14, 15].

Ambiguity is the most substantial concern in developing NER research. Some entity types in a sentence can be misclassified as another type of entity. For an example, there is some person name that is also a location name in Indonesian. We use a CRF as the baseline system, and it appears that some entity

has been misclassified. The features that is used in this baseline system are the surrounding word and syntactic features like Part-Of-Speech (POS).

However, in recent studies, there are some efforts in trying to obtain a word representation from unlabelled data. This word representation is known as word embedding. Word Embedding has an advantage, that it can capture semantic meaning of a words in document. Much research has been done to build word embedding, like Word2Vec in [16] and Glove in [17]. Numerous research in NER tries incorporating word embedding as a feature in the supervised classifier like in [18] and [19].

This study proposes a hybrid model to obtain Indonesian Named Entity by incorporating Word Embedding into baseline algorithms. Our hybrid model is built by combining the word embedding as a cluster features into Conditional Random Fields (CRF) like the previous study in [18]. We use K-Means as the clustering algorithm. The clustering results will be processed as an additional feature besides the standard contextual word and Part-of-Speech (POS) features into CRF.

Similar words in word embedding tend to have a similar vector. Therefore, these similar words will be grouped together in the same cluster. To capture this behavior, we use the cluster features in the CRF. It will help the CRF to obtain a better result, like a result of the previous study in [18]. So, it can be concluded that CRF and K-Means have complemented each other's strength in our proposed hybrid model for Named Entity Recognition. Word embedding in our research is pre-trained Word2Vec based on the previous study by [20].

Based on the methodology that we proposed in this study, we will describe several main contributions as follows:

- A hybrid CRF and K-Means clustering algorithm are proposed in this study.
- K-means algorithm is utilized to obtain the word cluster from Word2Vec, which significantly improves the performance of the algorithms by eliminating the ambiguity of misclassified type of entity from baseline algorithms.
- A standard dataset that can be used in Indonesian Named Entity Recognition task is rarely to be found. In this study, we try to propose this dataset as one of the standard datasets for Named Entity Recognition task in Indonesian. Our dataset was taken from Indonesian Online News from various topics.

This paper is divided into seven parts and organized as follows. The first part is the introduction in section 1. Section 2 describes the previous study in Named Entity Recognition. Section 3 describes how to obtain and construct the dataset for Indonesian Named Entity Recognition Dataset. Section 4 explains about our hybrid models, and section 5 will be discussing the experiments. Section 6 and 7 will be the last part of this study, that describes about discussion, conclusion, and further research.

2. Related works on named entity recognition

Named Entity Recognition (NER) or proper name classification is one of the main components in Information Extraction. NER is widely used to detect a person's name, location, and organization in a document. However, the entity type can be broadened in other kinds of entities according to the needs. As one of the vital research in Natural Language Processing, NER has become the foundation of other NLP tasks such as coreference resolution [21], machine translation [10, 22], and question answering system [23].

There are various methods to solve NER; one of them is rule-based NER by utilizing a data dictionary that consists of country, city, company, and name [24]. With the rule-based approach, entity recognition is conducted by defining rules about words position patterns in a sentence [25]. However, with rule-based and dictionary-based approach, there is a high dependency on the domain where the rules and dictionary were built. Thus the Named Entity Recognition will face challenges on a new domain or new sentence model.

Due to the previous limitations, machine learning approach is widely implemented in the Named Entity Recognition task. Hidden Markov Model (HMM) is one of the machine learning algorithms that is used beside Conditional Random Fields. NER with HMM does not need a language expert. This means it can be utilized in any language and could achieve a good performance score [26]. However, Conditional Random Fields is advantageous compared to HMM due to its method, which receives a sequence and maximizes conditional probabilities of labels so any additional feature could be easily represented. Another machine learning approach is by Maximum Entropy in [27], which is used to obtain Czech named entity. Nonetheless, the shortcoming of Maximum Entropy label bias problem could be covered by CRF.

NER has been implemented in various languages, such as Arabic [28], and especially in Indonesian. However, NER in Indonesian has similar challenges in English, though the language structure is different.

Table 1. Related works in Indonesian NER

No.	Authors	Tagset	Dataset	Model
1	Munarko et.al. [11]	Person, Location, Organization, Other	2000 training data from tweet	Conditional Random Fields
2	Jaariyah [29]	Person, Location, Organization, Other	2231 sentences	Conditional Random Fields
3	Wibawa et.al. [30]	Person, God, Organization, Location, Facility, Product, Event, Natural-Object, Disease, Color, Timex, Periodx, Numex, Countx, Measurement	457 news articles with 1500 sentences	Naive Bayes, SVM, Simple Logistic
4	Al-Ash et.al. [31]	Age, Date, Doctor, Hospital, ID, Location, Patient, Phone	888 documents	Combination of Long Short Term Memory and Conditional Random Fields
5	Wintaka et.al.[32]	Person, Location, Organization	600 Indonesian Tweets	Combination of Bidirectional Long Short Term Memory and Conditional Random Fields
6	Wibisono et.al. [33]	Person, Organization, Other	2092 sentences	Combination of Long Short Term Memory and Conditional Random Fields
7	Rachman et. al. [34]	Organization, Person, Location	480 tweets	BiDirectional Long Short Term Memory
8	Anggareska et. al. [35]	Object, Location, Time, Condition, Cause, Suggestion, Link	290 tweets	Naïve Bayes, SMO, IBK

When the NER task is done using a machine learning approach, the methods in English can also be implemented in Indonesian as in research [29]. The research still has an opportunity to be enhanced by selecting additional features.

Quite a lot of research on Named Entity Recognition has been proposed in Indonesian. We made a comparison of some research in Indonesian that can be seen in Table 1. A study by [30] suggested Ensemble Supervised Learning for 15 classes, which achieved the best performance with the Simple Logistic Algorithm at 52%. This accuracy score also could be improved by selecting better features.

Another example of Indonesian named entity recognition is done on Protected Health Information Removal in [31]. In this research, imbalanced data was a problem. Thus we provide a dataset with more annotation with a total of 51,241 named entity from 30,407 annotated sentences in this research.

In Anggareska et.al. [35], the best combination of features and algorithms for Named Entity Recognition and Relation Extraction was explored with three algorithms, which are Naïve Bayes, SMO, and IBK. The best accuracy was obtained with SMO. There were still errors which were due to ambiguity.

Deep Learning also has been explored as a model for Named Entity Recognition. A combination of Bidirectional LSTM and CRF was proposed in Indonesian [32]. However, Deep Learning has high computation costs. The previous research also didn't incorporate word embedding, which could have improved the performance. Our model achieved good results without that disadvantage. Research in Indonesian Named Entity Recognition was conducted in [11], without a hybrid model and solely depended on Conditional Random Field performance alone. The performance could be enhanced with the hybrid model developed in our research. Wibisono et.al.[33] performed Named Entity Recognition combining Bidirectional LSTM with CRF. This research would result in better performance by adding sufficient data in the NER Dataset.

3. Indonesian named entity recognition dataset

The dataset used in this research was taken from Indonesian online news with a total of 29,587 webpages taken from CNN Indonesia website. Each webpage consists of one news article. Each news article was crawled from the website using a crawler and performed a preprocessing using several processes that are displayed in Fig 1.

Web content extraction was done by using regular expressions and rules that are formed manually to extract the news content from the HTML on the webpage. Sentence boundary detection and tokenization are done with the help of the Spacy.io library with a pre-trained language model for Indonesian. The results of this process will then be stored as unlabeled data, which later be annotated as a Named Entity Recognition dataset in Indonesian.

Figure. 2 Dataset Example.

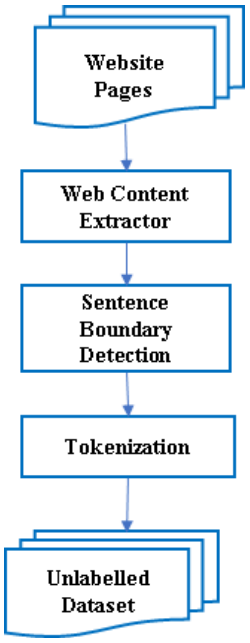

Figure. 1 Preprocessing for building Indonesian NER dataset

The labeling process is carried out by two annotators using Brat Annotation Tools [36]. The annotation is done by labeling part-of-speech and named entity label. The total part-of-speech tags in this dataset were 35 and obtained from the previous research in [37]. The entity labels have four types, namely person (PER), location (LOC), organization (ORG), and miscellaneous (MISC).

Person defines the entity for the name of the person. Location defines the entity for the location name. Organization is used to define the name of an organization. As for miscellaneous, it is used to represent entities other than the four types of those entities.

The agreement of the two annotators determines the label values in the final dataset. If there exists a difference in labels from two annotators, the sentence will not be saved as a final dataset. The final dataset available in this study consists of 30,407 sentences taken from approximately 2000 news documents. Examples of sentences used in this research are shown in Fig. 2.

There are 51,241 entities in our dataset. There are 25,817 person entities with a total of 50.39% from the entire dataset. The location entity takes about 12,088, which amounts to 23.59% of the whole dataset. The number of organization entity is the total of 9,881, or 19.28% of the total entities in the dataset. The miscellaneous types entity amounts to 3,455 with a percentage of 6.74% from all entities in the dataset. The final dataset will be converted from Brat annotation format into BIO for each entity, where:

a. B (Begin) marks the beginning of a word that is an entity.
b. I (Inside) marks a word that is part of the entity.
c. O (Outside) marks a word that is not an entity or part of an entity.

The number of labels contained in the document for classification is nine, which consists of eight entity labels and one O label. Eight labels include four types of entities that each of them represented by B and I. These nine labels will be guessed by the system to produce a set of Named Entities in the document.

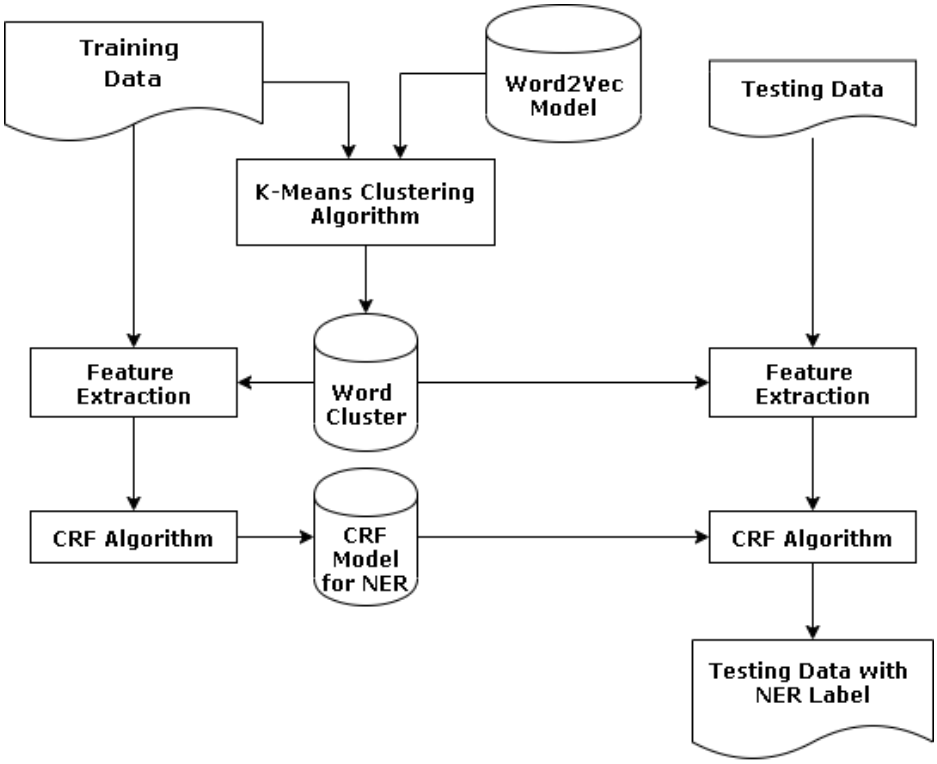

Figure. 3 Hybrid approach for named entity recognition

4. Hybrid conditional random fields and k-means for named entity recognition

This section will discuss the details of the proposed hybrid model. The architecture of the proposed hybrid model is found in Fig 3. The pseudocode of our proposed system is found in Algorithm 1. The training and testing data used in this study have been preprocessed using steps displayed in Fig. 1. Part-Of-Speech (POS) Tagging will be carried out after the preprocessing using tools from [37].

The algorithms take an input D_{train} which is a collection of training sentence *S* that consists of *word*, part-of speech (*POS*), and NER *label*. The first step is done by clustering each word in D_{train} using *W2V* model and store it into *w_cluster*. Each *word*, *POS*, and *label* in D_{train} alongside the *w_cluster* is used to construct the features for NER task. The features used by CRF to build the model and save it into *CRFModel*. This model will be used to extract the Named Entity in the D_{test}, which is the testing document contains the testing sentence *S* that consists of word and part-of-speech (POS). The named entity label for testing documents D_{test} is saved in Y_{test} and return as a result of the NER task. For a detail explanation of K-Means and CRF will be described in subsections 4.1 and 4.2.

Algorithm 1 ***Pseudocode for Hybrid NER***

Input : D_{train} : Training Dataset
D_{test} : Testing Dataset
W2V: Word2Vec Model
K : Number of Cluster in K-Means
Output : Y_{test} (NER Label for D_{test})

Pseudocode for Building the Model

```
w_cluster = K-Means(D_train, K, W2V)
features = []
target = []
FOR S IN D_train
    FOR word, pos, label IN S
        features.add(extractFeatures(word, pos,
                     w_cluster))
        target.add(label)
    NEXT
NEXT
CRF = initCRFModel()
CRFModel = trainUsingSGD(CRF,features, target)
RETURN w_cluster , CRFModel
```

Pseudocode for Named Entity Recognition

```
Y_test = []
FOR S IN D_test
    FOR word, pos IN S
        features = extractFeatures(word, pos,
                   w_cluster)
        label = CRFModel.predict(features)
        Y_test.add(label)
    NEXT
NEXT
RETURN Y_test
```

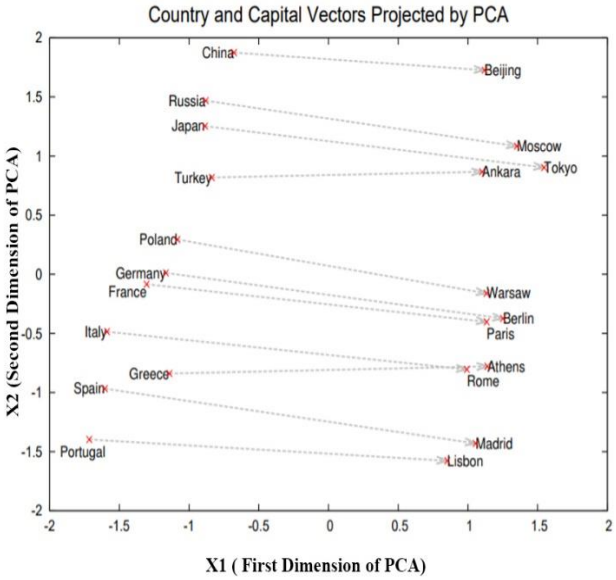

Figure. 4 Word2Vec illustration taken from Mikolov [38]

4.1 K-means algorithm

K-Means are used to form word clusters based on the closeness of words in Word Embedding. A group of words in a word cluster are words that are considered to have the same semantic meaning. It believes that the word cluster can improve the performance of Conditional Random Fields.

The type of word embedding used in this study is Word2Vec, which was formed using the SkipGram Negative Sampling model [38]. Word2vec has several advantages, including being able to describe the semantic closeness between words, which can be seen in Fig 4.

Each word in Fig. 4 is taken from SkipGram model for English with a dimension size of 1000. These words vector dimension is simplified from 1000 to 2, so it can be visualized into 2D space using PCA Algorithms. The X axis in Fig. 4 indicates the first dimension of PCA results, and the Y axis in Fig. 4 indicates the second dimension of PCA results for each word.

We carried out some example in Indonesian Word2Vec by giving the word based on entity types such as person (PER), location (LOC), organization (ORG), and miscellaneous (MISC). We try to get the top 5 closest words found in the word2vec model. It is proving that those words have the same type of entity. Details of the results is displayed in Table 2.

As an example, from Table 2, given the input word Joko, the set of words returned is a row of names of people, including Widodo, Bambang, etc. For other types such as entity type MISC, given a disease name, then the 10 closest words are several types of diseases and viruses, such as coronavirus, zika, and gastroenteritis.

Table 2. Top 5 nearest words in Indonesian language Word2Vec

No.	Words	Entity Type	Nearest Words
1	Joko	PER	Widodo, Bambang, Susilo, Jokowi, Wiranto
2	Surabaya	LOC	Malang, Semarang, Jember, Sidoarjo, Madiun
3	LIPI	ORG	LEKNAS, PUSLIT GFTK, BPPT, PUSLITBANG
4	SARS	MISC	Coronavirus, Kaposi, Zika, Crohn, Gastroenteritis

It has been concluded that similar words with the same meaning tend to be clustered together. In our proposed hybrid model, the word cluster is built to help CRF improve its performance. Clusters size in this research is taken from size $k = 100$ sd. $k = 500$. A set of words in the training data is used as an input to the K-Means Algorithm. The clustering process is done by using the vector value from each word in the Word2Vec model.

4.2 Conditional random fields

This section will describe the Conditional Random Fields (CRF) algorithm for Named Entity Recognition. Conditional Random Fields is a statistical probabilistic graphical model that has been utilized for segmenting and labeling sequence data and shows an advantage over Hidden Markov Models [39].

Given a set of words w_1, w_2, w_3, …, w_n with POS Tag p_1,p_2,p_3,…,p_n in sentence S with length n, then it will be used as features to guess the correct entity label of word w_i.. The CRF algorithm uses contextual features from around the word w_i that was proposed at [40], [41]. Other features added in this our proposed methodology are prefix+suffix and word cluster. Explanation of the features was used in this proposed methodology are:

a. Words Features
   Words appearing around w_i words are believed can help determine the type of entity. The word features of Eq. (1) will be used to predict the label of the word w_i.

$$word(w_i) = [w_{(i-j)}, \ldots, w_i, \ldots, w_{(i+j)}] \quad (1)$$

where w_i is the *i-th* word in the sentence, and *j* is the window size for contextual features.

b. Part-Of-Speech (POS) Feature
Part-of-Speech is a grammatical tagging of the word w_i in the sentence. The use of POS as a feature in NER is very important because it can help the model determine the type of entity according to the context in the sentence. The POS tag feature used can be seen in Eq. (2).

$$pos(w_i) = [p_{(i-j)}, \dots, p_i, \dots, p_{(i+j)}] \quad (2)$$

where p_i is the POS tag of the word *i-th* in sentence, and *j* is the window size of contextual features.

c. Prefix + Suffix Feature
In addition to using word features, this research adds prefixes and suffixes of the words as features. This technique is used to overcome Out of Vocabulary (OOV) words that never appear in the training data.

A collection of words that have the same prefix and suffix are believed to have the same type of entity. Prefixes and suffixes of each word feature will be taken from one to three characters. The prefix and suffix features can be obtained using Eq. (3) and Eq. (4).

$$\begin{aligned} prefix(w_i) = [&npreff(w_{(i-j)},\dots,w_{(i+j)},1),\\ &npreff(w_{(i-j)},\dots,w_{(i+j)},2)\\ &npreff(w_{(i-j)},\dots,w_{(i+j)},3)] \end{aligned} \quad (3)$$

$$\begin{aligned} suffix(w_i) = [&nsuff(w_{(i-j)},\dots,w_{(i+j)},1),\\ &nsuff(w_{(i-j)},\dots,w_{(i+j)},2),\\ &nsuff(w_{(i-j)},\dots,w_{(i+j)},3)] \end{aligned} \quad (4)$$

where *npreff(w, n)* is a function to get the first *n* characters of the word *w*, while *nsuff(w, n)* is a function to get the last *n* characters of the word *w*. The *n* value of the *nsuff* and *npreff* functions is one to three. Parameter *w* in *npreff* and *nsuff* function is the word being sought for its prefix or the suffix. Variable *i* defines the position of the word in the sentence, and *j* is the window size of contextual features.

d. Word Cluster Feature
This feature is the clustering result from the K-Means algorithm. A cluster label for each word will be used as features. The equation for obtaining this feature is in Eq. (5).

$$\begin{aligned} word\_cluster(w_i) = [&wc(w_{(i-j)}), \dots, wc(w_i),\\ &\dots, wc(w_{(i+j)})] \end{aligned} \quad (5)$$

Where *wc(w)* is a function to get the cluster label for a word *w* from the k-means cluster results. The variable *i* shows the position of the word *w* in the sentence and *j* is the window size for contextual features.

Conditional Random Fields (CRF) is a probabilistic statistical model that is commonly used to label sequence data [39]. One of the CRF models is that is widely used in Natural Language Processing is Linear-chain CRF. The features used in this research will be converted into a feature function based on the example in Eq. (6). Y_i is the label for word *i-th* in the sentence. Y_{i-1} is the label of the word $w_{(i-1)}$ or the previous word, and x_i is the feature for the word *i-th*. As an example, in Eq. (6), the word that appear at the *i-th* position in the sentence is *Surabaya*. Therefore the type of entity result is B-LOC, since Surabaya is the capital city of East Java province.

$$f(y_i,y_{i-1},x_i) = \begin{cases} 1, if\ y_i = B-LOC\ and\ y_{i-1} = O \\ \quad and\ x_i = Surabaya \\ 0, else \end{cases} \quad (6)$$

To perform classification, the CRF algorithm will look for maximum conditional probability $P(Y|X)$ where *Y* is the sequence of labels and *X* is the sequence of words. The conditional probability model $P(Y|X)$ can be seen in Eq. (7) and Eq. (8).

$$P(Y|X) = \frac{1}{Z(X)} \prod_{j=1}^{N} exp \sum_{i=1}^{M} \lambda_i f(y_i, y_{i-1}, x_j) \quad (7)$$

$$Z(X) = \sum_{i=1}^{Yn} \prod_{j=1}^{N} exp \sum_{i=1}^{M} \lambda_i f(y_i, y_{i-1}, x_j) \quad (8)$$

Where *N* is the number of sequences, *M* is the number of features, λ_i is the parameter for the *i-th* feature, and Y_n contains the number of labels that will be recognized in the CRF. Parameter λ will be optimized using the Stochastic Gradient Descent (SGD) algorithm by minimizing the loss function based on Eq. (9).

$$L(\lambda, D) = -log\left(\prod_{k=1}^{Yn} P\left(y^k|x^k,\lambda\right)\right) + C\frac{1}{2}||\lambda||^2 \quad (9)$$

On Eq. (9), variable *D* is training data where $D = [(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N)]$. The variable *x* in *D* is the sequence of words, and *y* is the sequence of labels. This training data *D* has the *N* sequence length.

The variable *C* in Eq. (9) define the L2 Regularization parameter for the loss function calculation.

5. Experiments

This section will discuss how experiments will be carried out. The results of each experiment will be analyzed to see the performance of the proposed model.

5.1 Experiments scenario

In this section, we will discuss our experiment scenario that will be done. Word2vec used in this study is Skip-Gram, which was formed with Negative Sampling. The model was built using Indonesian Wikipedia articles with a total of 308,227 articles. The experiments will be divided into three scenarios which are:

- Baseline(B): Experiment using standard features, which are contextual window word features and part-of-speech.
- Baseline+PrefixSuffix(B+PS): The experiment is carried out by adding Prefix and Suffix features in the Contextual Window Feature.
- Baseline+PrefixSuffix+W2V(B+PS+W2V):This experiment is done by adding the word cluster features obtained from the K-Means algorithm. In this experiment, the number of k used was 100 until 500.

The performance of our proposed model will be evaluated using F1-Score used in CoNLL 2003[42], which can be seen in Eq. (10) to Eq. (12).

$$F1 - Score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (10)$$

$$Precision = \frac{total\ correct\ entity\ from\ model}{total\ entity\ from\ model} \quad (11)$$

$$Recall = \frac{total\ correct\ entity\ from\ model}{total\ gold\ entity\ in\ dataset} \quad (12)$$

The experiment will be carried out by dividing the amount of training and testing data using percentage splitting. The total amount of data used was 30,407 sentences with four types of entities, namely person (PER), organization (ORG), location (LOC), and miscellaneous (MISC). The experiment will be divided into 4 types, namely by dividing 60-40, 70-30, 80-20, and 90-10.

Other experiments were done to compare the proposed model in this study with Bidirectional Long Short Term Memory (Bi-LSTM) in [34], which is a standard model in sequential tagging and also

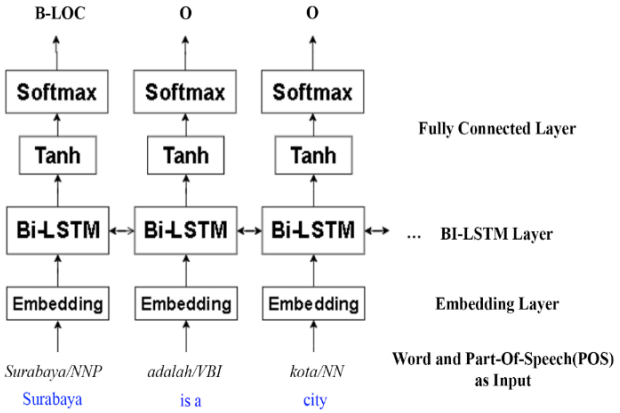

Figure. 5 Comparable models

BILSTM-CRF that already widely used in nowadays sequential tagging framework in [32, 33]. We have conducted experiments with these previous researches with our dataset. An illustration of the model compared in this study can be seen in Fig. 5.

The Bi-LSTM was compared using two scenarios, like in the previous research [34]. The first scenario takes an input of word, and part-of-speech (POS) called BILSTM+WE+POS. Whereas in the second scenario, we use words only as an input called BILSTM+WE. The model uses pre-trained embedding Word2Vec to represent words while embedding for part-of-speech will be trained together with the model during learning. And for the third scenario for comparison with the BILSTM-CRF, we used the same scenario like in [33].

5.2 Experiments results

The first experiment was carried out using the data-training and testing division of 60-40. The total training sentence used in this experiment is 18,245 sentences, and the testing sentence is 12,162. The results of the experiment can be seen in Table 3.

The second experiment was carried out by dividing the whole sentence into 70-30. The total training sentence used was 21,284, while for the testing sentence, it was 9,123. The results of this experiment can be seen in Table 4.

The third experiment was carried out with the data distribution of 80-20. The total sentence used as training data in this experiment was 24,325. For the entire testing sentence used amounted to 6,082. The results of the third experiment can be seen in Table 5.

The fourth experiment was carried out by dividing the data into 90-10. The sentences used as training data are 27,366. Whereas for testing, there were 3.041. The results of the experiments obtained can be seen in Table 6.

Table 3. Experiments on 60-40 data distribution

No	Model	PER	ORG	LOC	MISC	ALL
1	B	87.06	76.52	84.17	58.92	82.31
2	B+PS	90.64	82.74	86.98	63.41	86.24
3	B+PS +W2V (100)	91.92	82.64	88.21	62.93	87.12
4	B+PS +W2V (200)	92.03	82.79	88.01	63.25	87.18
5	B+PS +W2V (300)	91.69	82.96	88.03	62.23	86.97
6	B+PS +W2V (400)	91.58	82.23	88.56	61.44	86.84
7	B+PS +W2V (500)	91.71	83.02	88.6	63.07	87.18

Table 4. Experiments on 70-30 data distribution

No	Model	PER	ORG	LOC	MISC	ALL
1	B	86.45	76.37	83.77	44.62	81.29
2	B+PS	89.91	82.79	87.52	48.2	85.38
3	B+PS +W2V (100)	91.22	81.93	87.63	48.21	85.89
4	B+PS +W2V (200)	91.55	82.73	87.89	48.43	86.3
5	B+PS +W2V (300)	91.09	82.26	87.79	48.82	85.97
6	B+PS +W2V (400)	91.26	82.33	88	48.01	86.06
7	B+PS +W2V (500)	91.08	82.35	88.22	48.07	86.03

Table 5. Experiments on 80-20 data distribution

No	Model	PER	ORG	LOC	MISC	ALL
1	B	85.52	79.3	82.78	46.22	80.88
2	B+PS	89.3	84.02	86.37	52.81	84.92
3	B+PS +W2V (100)	90.41	82.84	86.65	54.64	85.3
4	B+PS +W2V (200)	90.67	83.61	86.86	52.54	85.57
5	B+PS +W2V (300)	90.92	83.62	86.53	52.75	85.64
6	B+PS +W2V (400)	90.84	83.98	86.66	52.86	85.71
7	B+PS +W2V (500)	90.64	83.97	87.03	52.25	85.67

Table 6. Experiments on 90-10 data distribution

No	Model	PER	ORG	LOC	MISC	ALL
1	B	87.06	76.52	84.17	58.92	82.31
2	B+PS	88.12	84.05	85.77	52.69	84.13
3	B+PS +W2V (100)	88.83	83.54	86.74	52.48	84.54
4	B+PS +W2V (200)	89.01	84.53	86.72	52.63	84.89
5	B+PS +W2V (300)	89.53	85.13	86.17	51.36	85.1
6	B+PS +W2V (400)	89.07	84.6	86.68	52.06	84.88
7	B+PS +W2V (500)	88.8	84.71	86.78	51.71	84.78

The fifth experiment was carried out by comparing with other models from previous research. There are three scenario what already mention in the subsection 5.1. Some experimental results can be seen in Table 7.

Detailed explanation and discussion of experiment results will be explained in section 6. According to the experiment results, our proposed model can achieve better performance compared with the baseline and other models.

6. Discussion

Our experiments are done using 2 types of scenarios. First is using percentage splitting, and the second scenario is done by comparing our methods with previous research. Each experiment in percentage splitting was done using several scenarios that describe in subsection 5.1. Each experiment was conducted by a combination of from each feature that are proposed in this paper.

The first experiment was done by using percentage splitting on 60-40. The result is shown in Table 3. The baseline model only achieved 82.31% performance. The combination of baseline and prefix suffix for the OOV words can improve the baseline into 86.24%. The best performance was obtained from two models with cluster size 200 and 500.

Experiments with 70-30 data splitting in Table 4 gave the best results of 86.3% using hybrid CRF and K-Means models with 200 number of clusters. When

Table 7. Performance comparison of our proposed methods with other models

No	Model	PER	ORG	LOC	MISC	ALL
1	BI-LSTM +WE +POS	89.93	76.59	81.62	59.48	83.09
2	BI-LSTM +WE	89.33	76.29	80.91	61.10	82.77
3	BI-LSTM +CRF	88.32	76.92	84.93	62.96	83.10
4	Our Method	92.03	82.79	88.01	63.25	87.18

compared to the baseline model, the performance of the proposed hybrid model increased by 5.01%. On the other hand, the performance of the baseline model with prefix and suffix alone gives 84.92% performance and has a 1.38% difference compared to the best model obtained in this study.

Table 5 shows experiments with 80-20 percentage splitting, the best performance was obtained 85.71% of the hybrid model with the number of clusters of 400. When compared to the baseline model, the best performance of the hybrid model increased by 4.83%. Meanwhile, when compared with experiments with baseline and suffix prefixes, hybrid models provide an increase of 0.79%.

The final experiment for percentage splitting is diving by 90-10 on Table 6. The best results obtained from the hybrid model with the number of clusters is 300, which is 85.1%. Experiments with baseline models obtained an F1-Score of 82.31%. As for the results of experiments with baseline and suffix prefixes, we obtained F1-Score of 82.13%. A performance increase of 2.97% was obtained from the hybrid model compared to the baseline.

The experimental results in percentage splitting scenario shows that the best performance is obtained with the proposed model with the combination of Word2Vec cluster features with k = 200. Clusters are believed to improve the performance of the model by grouping words that have similar meanings. The results of using the k value for the K-Means algorithm do not give a significant difference in performance. However, the use of hybrid models is proven to provide improved performance compared to several other models.

Another contribution offered in this study besides the use of Hybrid model, is the use of prefix and suffix of a word as a feature of the CRF algorithm. The prefix and suffix of a word are believed to help the model to determine the type of entity. The same type of entity will tend to have the same prefix and suffix and can help solve words that are OOV. The addition of the suffix prefix feature has increased the impact from the proposed baseline model. Proof of increasing the results can be seen in Tables 3, 4, 5, and 6.

In addition to percentage splitting, other experiments were carried out by comparing the proposed methodology with the previous research. We make a comparison with two models that are widely used in NER. First is we compare with the BILSTM model and BILSTM-CRF. The BILSTM was trained using two types of combination input, namely word, and POS.

The result of our proposed method can give a better performance compared to this model. Our model can achieve better performance by 4.09% compare to the BISTLM that uses Word Embedding and POS Embedding. The second comparison was made by taking only the Word Embedding as an input to the BILSTM. The result shows that our model can achieve better performance by 4.41%. Another type of model was used as a second model. This model is BILSTM-CRF that taken from previous research. BILSTM-CRF was one of the current state-of-the art algorithms in sequential neural tagging. The results of the proposed model in this study reveal a performance increase of 4.7% for BILSTM-CRF.

When we look to the result of NER, the worst performance is achieved by miscellaneous entity type, and the best performance was taken from person entity type. The combination of entity types that are categorized as miscellaneous causes ambiguity. The ambiguity in the miscellaneous entity type in this recognition causes the performance of the model to be unstable and tends to fail to be recognized.

Based on the experiments, the performance of the proposed model can provide a fairly good performance. Thus, the proposed method is expected to be one of the state-of-the-art for NER in Indonesian.

7. Conclusions and further research

From the results of the experiments performed, the proposed hybrid model has the best result of 87.18% with the training and testing data splitting of 60-40. The performance of the methods offered can provide the best improvement on average by 4.37% from all the experiments. When we compare the model with other models, we can get better performance on average by 4.25%. The hybrid model achieved better results compared to the baseline single model.

The word clusters in this research greatly affected the performance of the CRF, even though the value

of the F1-Score didn't change significantly. The best result from all experiments is obtained with parameter k=200 of the K-Means Algorithm. In future works, we need to determine the optimal number of clusters in order to obtain maximum results.

The experiments that have been conducted prove an excellent performance in NER task for Indonesian. The comparison with the previous works shows that our model can achieve better performance with averaging 4.3% compared to the existing techniques with the Deep Learning approach. The comparable models also require a lot of data and had to handle imbalanced data.

Our proposed method is a combination of two models which are CRF as the state-of-the-art algorithm for Named Entity Recognition that is widely used as the state-of-the art approach before the Deep Learning Era. Additionally, word representation is done by applying word embedding in current research. However, it is a challenging task to integrate the continuous vector to the graphical model like Linear Chain CRF.

Thus, with our proposed model, we used K-Means to make word clusters as a feature that could be incorporated into CRF. Based on our experiments in this paper, our proposed method worked well compared to the standard CRF that is widely used in Named Entity Recognition. However, for future research, a comparison can be made for other word embedding methods, such as fastText and Glove.

Acknowledgments

We would like to say thanks to friends from ITS for their comments and critique while writing this article. The authors also would like to express gratitude to Dr. Gunawan, Christian Nathaniel Purwanto, Amelinda Tjandra Dewi, and the members of the Computational Linguistics Research Group in ISTTS for the help to prepare the initial data.

References

[1] S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, and B.Xu, "Joint entity and relation extraction based on a hybrid neural network", *Neurocomputing*, Vol. 257, pp. 59–66, 2017.

[2] K. Yang, Y. Cai, D. Huang, J. Li, Z. Zhou, and X. Lei, "An effective hybrid model for opinion mining and sentiment analysis", In: *Proc. of 2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 465–466, 2017.

[3] H. Akkineni, V. S. L. Papineni, and V. B. Burra, "Hybrid method for framing abstractive summaries of tweets", *International Journal of Intelligent Engineering and Systems*, Vol. 10, No. 3, pp. 418-425, 2017.

[4] P. Katta and N. P. Hegde, "A Hybrid Adaptive Neuro-Fuzzy Interface and Support Vector Machine Based Sentiment Analysis on Political Twitter Data", *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 1, pp. 165–173, 2019.

[5] K. Xu, Z. Zhou, T. Gong, T. Hao, and W. Liu, "SBLC: a hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields", *BMC Medical Informatics and Decision Making*, Vol. 18, No. 5, pp. 114, 2018.

[6] T. H. Pham and P. Le-Hong, "End-to-End Recurrent Neural Network Models for Vietnamese Named Entity Recognition: Word-Level Vs. Character-Level", *Communications in Computer and Information Science*, Vol. 781, pp. 219-232, 2018.

[7] G. Aguilar, S. Maharjan, A. P. López-Monroy, and T. Solorio, "A multi-task approach for named entity recognition in social media data", *In: Proc. of the 3rd Workshop on Noisy User-generated Text*, pp. 148-153, 2017.

[8] H. S. Al-Ash and W. C. Wibowo, "Fake News Identification Characteristics Using Named Entity Recognition and Phrase Detection", In: *Proc. of 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 12–17, 2018.

[9] F. Nurifan, R. Sarno, and K. R. Sungkono, "Aspect Based Sentiment Analysis for Restaurant Reviews Using Hybrid ELMo-Wikipedia and Hybrid Expanded Opinion Lexicon-SentiCircle", *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 6, pp. 47–58, 2019.

[10] N. Agrawal and A. Singla, *Using named entity recognition to improve machine translation*, Stanford University Natural Language Processing, San Fracisco, S.F. 2012.

[11] Y. Munarko, M. S. Sutrisno, W. A. I. Mahardika, I. Nuryasin, and Y. Azhar, "Named entity recognition model for indonesian tweet using crf classifier", *IOP Conference Series: Materials Science and Engineering*, Vol. 403, No. 1, pp. 1-6, 2018.

[12] F. Souza, R. Nogueira, and R. Lotufo, "Portuguese Named Entity Recognition using BERT-CRF", *arXiv Prepr. arXiv1909.10649*, pp. 1-8, 2019.

[13] H. Gasmi, A. Bouras, and J. Laval, "LSTM recurrent neural networks for cybersecurity

named entity recognition", In: *Proc. of The Thirteenth International Conference on Software Engineering Advances*, pp. 1-6, 2018.

[14] B. Y. Lin, F. F. Xu, Z. Luo, and K. Zhu, "Multi-channel bilstm-crf model for emerging named entity recognition in social media", In: *Proc. of the 3rd Workshop on Noisy User-generated Text*, pp. 160–165, 2017.

[15] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition", *Bioinformatics*, Vol. 34, No. 8, pp. 1381–1388, 2018.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", In: *Proc. of ICLR Workshop*, Arizona, USA, pp. 1-12 ,2013.

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation", In: *Proc. of 2014 Conference of Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[18] V. Suárez-Paniagua, I. Segura-Bedmar, and P. Mart'inez, "Word embedding clustering for disease named entity recognition", In: *Proc. of the fifth BioCreative Challenge Evaluation Workshop*, pp. 299–304, 2015.

[19] M. Seok, H.J. Song, C. Park, J.D. Kim, and Y.S. Kim, "Named Entity Recognition using Word Embedding as a Feature", *International Journal of Software Enginerring and Its Applications*, Vol. 10, No. 2, pp. 93–104, 2016.

[20] J. Santoso, A.D.B. Soetiono, Gunawan, E. Setyati, E.M. Yuniarno, M. Hariadi, and M.H. Purnomo, "Self-Training Naive Bayes Berbasis Word2Vec untuk Kategorisasi Berita Bahasa Indonesia", *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, Vol. 7, No. 2, pp. 158–166, 2018.

[21] Z. Dai, H. Fei, and P. Li, "Coreference aware representation learning for neural named entity recognition", In: *Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4946-4953, 2019.

[22] A. Jain, B. Paranjape, and Z. C. Lipton, "Entity Projection via Machine-Translation for Cross-Lingual NER", In: *Proc. of 2019 Conference of Empirical Methods in Natural Language Processing*, pp.1083-1092, 2019.

[23] A. Lamurias and F. M. Couto, "Lasigebiotm at mediqa 2019: Biomedical question answering using bidirectional transformers and named entity recognition", In: *Proc. of the 18th BioNLP Workshop and Shared Task*, pp. 523–527, 2019.

[24] R. Grishman, "The NYU System for MUC-6 or Where's the Syntax?", In: *Proc. of the Sixth Message Understanding Conference (MUC-6)*, pp. 167-175, 1995.

[25] T. Eftimov, B. K. Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations", *PLoS One*, Vol. 12, No. 6, pp. 1-32, 2017.

[26] S. Morwal, N. Jahan, and D. Chopra, "Named entity recognition using hidden Markov model (HMM)", *International Journal in Natural Language Computing*, Vol.1, No.4, pp. 15–23, 2012.

[27] M. Konkol and M. Konopík, "Maximum Entropy Named Entity Recognition for Czech Language", In: *Proc. of the 14th international conference on Text, Speech and Dialogue*, pp. 203–210, 2011.

[28] I. El Bazi and N. Laachfoubi, "Arabic Named Entity Recognition Using Topic Modeling", *International Journal of Intelligent Engineering and Systems*, Vol. 11, No. 1, pp. 229-238, 2017.

[29] N. Jaariyah, *Pengenalan Entitas Bernama Pada Teks Bahasa Indonesia Menggunakan Conditional Random Fields*, Universitas Komputer Indonesia, Indonesia, I.D. 2017.

[30] A. S. Wibawa and A. Purwarianti, "Indonesian named-entity recognition for 15 classes using ensemble supervised learning", *Procedia Computer Science*, Vol. 81, pp. 221–228, 2016.

[31] H. S. Al-Ash, I. Fanany, and A. Bustamam, "Indonesian Protected Health Information Removal using Named Entity Recognition", In: *Proc. of 2019 12th International Conference on Information & Communication Technology and System*, pp. 258–263, 2019.

[32] 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.

[33] Y. Wibisono and M. L. Khodra, "Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin", In: *Proc. of Seminar Tahunan Linguistik 2018*, pp. 1-5, 2018.

[34] V. Rachman, S. Savitri, F. Augustianti, and R. Mahendra, "Named entity recognition on Indonesian Twitter posts using long short-term memory networks", In: *Proc. of 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp.228–232, 2017.

[35] D. Anggareska and A. Purwarianti, "Information extraction of public complaints on Twitter text for bandung government", In: *Proc. of 2014 International Conference on Data and Software Engineering*, 2014.

[36] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: a web-based tool for NLP-assisted text annotation", In: *Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107, 2012.

[37] A. F. Wicaksono and A. Purwarianti, "HMM based part-of-speech tagger for Bahasa Indonesia", In: *Proc. of Fourth International MALINDO Workshop*, pp. 1-7, 2010.

[38] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", In: *Proc. of the 26th International Conference on Neural Information Processing Systems*, pp. 3111–3119, 2013.

[39] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In: *Proc. of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282–289, 2001.

[40] M. Tkachenko and A. Simanovsky, "Named entity recognition: Exploring features", In: *Proc. of KONVENS*, pp.118–127, 2012.

[41] J. Santoso, G. Gunawan, H. V. Gani, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Noun phrases extraction using shallow parsing with C4.5 decision tree algorithm for Indonesian Language ontology building", In: *Proc. of 2015 15th International Symposium on Communications and Information Technologies, ISCIT 2015*, pp.149-152, 2015.

[42] E. F. T. K. S. Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition", In: *Proc. of the Seventh Conference on Natural Language Learning at HLT-NAACL*, pp. 142-147, 2003.