

PAPER • OPEN ACCESS

## 5W1H Information Extraction with CNN-Bidirectional LSTM

To cite this article: A Nurdin and N U Maulidevi 2018 *J. Phys.: Conf. Ser.* **978** 012078

View the [article online](#) for updates and enhancements.



### 240th ECS Meeting

Oct 10-14, 2021, Orlando, Florida

**Register early and save  
up to 20% on registration costs**

Early registration deadline Sep 13

**REGISTER NOW**



# 5W1H Information Extraction with CNN-Bidirectional LSTM

A Nurdin<sup>1</sup> and N U Maulidevi<sup>2</sup>

<sup>1,2</sup>School of Electrical Engineering and Informatics, Institut Teknologi Bandung,  
Bandung, Indonesia

Email : <sup>1</sup>23515035@std.stei.itb.ac.id, <sup>2</sup>ulfa@stei.itb.ac.id

**Abstract.** In this work, information about *who, did what, when, where, why, and how* on Indonesian news articles were extracted by combining Convolutional Neural Network and Bidirectional Long Short-Term Memory. Convolutional Neural Network can learn semantically meaningful representations of sentences. Bidirectional LSTM can analyze the relations among words in the sequence. We also use word embedding word2vec for word representation. By combining these algorithms, we obtained F-measure 0.808. Our experiments show that CNN-BLSTM outperforms other shallow methods, namely IBk, C4.5, and Naïve Bayes with the F-measure 0.655, 0.645, and 0.595, respectively.

**Keywords:** information extraction; 5w1h; deep learning; Convolutional Neural Network; Bidirectional Long Short-Term Memory

## 1. Introduction

News articles provide information about event that consist of information *who did what, when, where, why, and how* (5W1H). Unfortunately, people often miss these main information since there are too much information about an issue, known as information overload. To overcome this problem, 5W1H extraction system can automatically present the main information in each news is proposed.

Information extraction is a task in Natural Language Processing (NLP) that aims to extract semantic content from a text [1]. This task can be solved by deep learning approach. In recent years, deep learning in machine learning field have succeeded show the best performance in classification task e.g. speech recognition [2], multimedia classification [3], and natural language processing [4-12]. To the best of our knowledge, there is no existing 5W1H information extraction research on Indonesian news articles using deep learning method.

In NLP task, convolutional neural network have been proven to be efficient for capturing meaningful representation of sentences e.g. classification and language modeling [4], sentiment analysis [5], named tagging and semantic role labeling [6], and event extraction [7,8]. Bidirectional LSTM is a type of RNN that popular for sequential modeling and achieving the best performance in machine translation [9], part-of-speech tagging [10], and language modeling [11,12].

In this paper, the task of 5W1H extraction was transformed into text categorization. Similar to previous research in 5W1H extraction on Indonesian news article [13-15], this task is performed as sequence tagging with BIO (Begin In Other) labelling scheme. Each token in an article will be classified into one of 13 classes. Label token depends on other token observations in the sequence and information of the entire sentence. We propose method that using CNN to extract information syntactic and semantic token within the sentence and using bidirectional LSTM to learn the relations among tokens in the sequence.



The rest of the paper is organized as follows. In section 2, related works in this area are presented. Section 3 describes our method. In section 4, experiments are discussed. Section 5 presents conclusion and future work.

## 2. Related Work

The research to extract 5W1H information on Indonesian news articles has been done by Khodra [13], Wicaksana [14], and Ilyas et.al. [15]. The three studies used machine learning to get key information in the news, what events occurred, the participants involved, the time and the location of the incident, why and how the incident occurred.

Khodra [13] extracted 5W1H information on 90 news articles using the C4.5 and AdaboostM1 algorithms with features set consist of the currentWord, bef1Word, bef2Word, currentTag, bef1Tag, bef2Tag, currentNETag, bef1NETag, bef2NETag, bef1Label, Bef2Label, isTimePattern, and isTitleToken.

Another research was conducted by Ilyas et.al [15]. In their research, sequence labeling technique, such as previous research, was used on 190 news articles and overcame the imbalanced dataset by using SMOTE (Synthetic Minority Over-Sampling). IBk (Instance-based Learning) algorithms, SVM, and C4.5 were used for classifying. The features used in their research are token, token kind, contextual, morphological, post, ne, sentence\_number, location, bef-one-token, bef-two-token, bef-one-post, bef-two-post, bef-one-ne, bef-two-ne, bef-one-class, bef-two-class, and current class.

Wicaksana [14] also conducted research to extract 5W1H information. The dataset used consisted of 80 news articles. Wicaksana [14] used bef1class, idxsentence, bef1pofe, and ne. The algorithms which used in the study are Naïve Bayes and IBk.

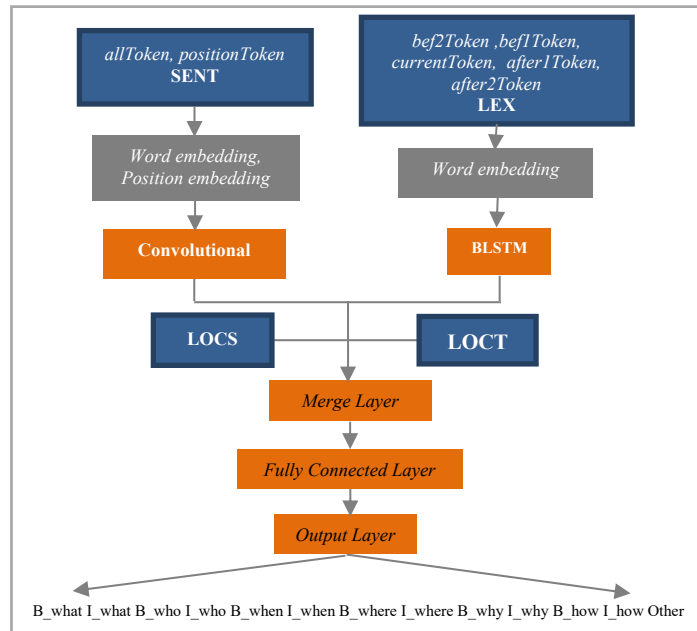
Unlike the previous researches [13-15] that used shallow learning algorithm, this research use deep learning CNN and BLSTM algorithm for modeling 5W1H information extraction in news articles. In addition for representation word, we use word2vec [16] which can represent the semantic distribution of words. In Khodra [13], Wicaksana [14], and Ilyas [15] research, the performance of the algorithm was evaluated by using 10-fold cross validation. The use of cross validation in Ilyas [15] and Wicaksana [14] was performed after resampling the dataset with the oversampling method of SMOTE. This could lead to overfitting as there may be replication of training data appearing in fold testing. Therefore, there is the same training data as the testing data. To avoid this phenomenon, cross validation is performed to split data into training and testing set then apply oversampling only on training set or by providing testing data for evaluation. In this research, we divide the dataset into training and testing data. Random oversampling is applied to train data to overcome the imbalanced dataset.

Chen, et.al. [7] and Nguyen, et.al. [8] employed CNN for event detection on ACE (Automatic Content Extraction) 2005 corpus with 529 training documents and 40 testing documents. Nguyen, et.al. [8] used CNN to detect events by transforming each token into real-valued vector via word embedding, position embedding, and entity type embedding. Chen, et.al. [7] used a CNN with dynamic multi-pooling to extract sentence and lexical level features so that it can detect trigger and argument of events in a sentence. To extract sentence level features, three types of inputs are used i.e. context-word features, position features, and event-type features. All inputs are transformed into real-valued vectors using word embedding, position embedding, and event type embedding and then combined into an input matrix on the convolutional layer. The output from the pooling layer is combined with lexical level features of trigger and argument candidate, one token before and one token afterwards.

## 3. Methods

We formalize the 5W1H information extraction as a multi-class classification problem. For each token in an article, we want to predict if the current token is an element of *what*, *who*, *where*, *when*, *why*, or *how*. We use BIO notation, as sequence tagging labelling scheme. Each token will be classified into

one of 13 classes (*B\_who*, *I\_who*, *B\_what*, *I\_what*, *B\_when*, *I\_when*, *B\_where*, *I\_where*, *B\_why*, *I\_why*, *B\_how*, *I\_how*, *Other*).



**Figure 1.** Proposed Network Architecture

Label of each token depends on other token observations in the sequence. Here, sequence tokens consist of current token and context token (token before and after current token). For example, consider the following sentences:

**S1:** **Indonesia** menjadi tuan rumah penyelenggara SEA GAMES XXVI.  
(*Indonesia* hosts The 26<sup>th</sup> SEA GAMES.)

**S2:** SEA GAMES XXVI akan diselenggarakan di **Indonesia**.  
(The 26<sup>th</sup> SEA Games will be held in *Indonesia*.)

In S1, *Indonesia* is a subject that labeled as *who* element. In S2, *Indonesia* indicates a location name, so that labeled as *where* element. This label depends on the context token. In S1, *Indonesia* is followed by a verb, *menjadi*. *Indonesia* tells the subject of the event. As a general rule, a subject word usually followed by a predicate word or a verb. In S2, *Indonesia* is a location because it is preceded by preposition of place and position, *di*.

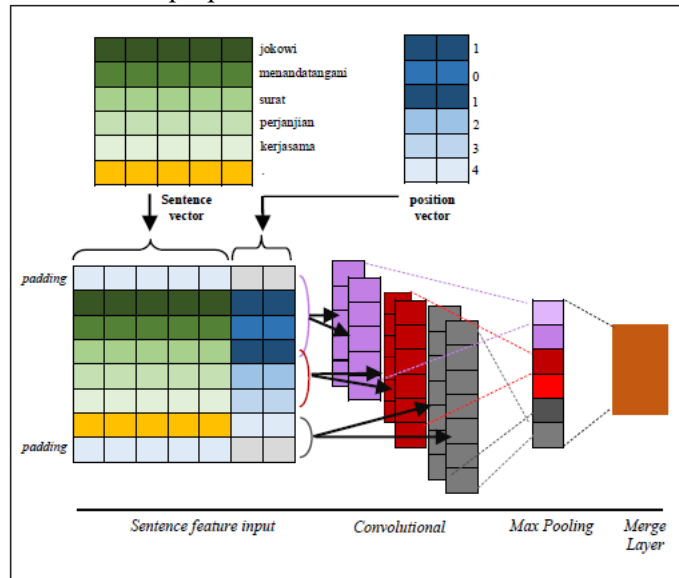
Beside information about current token and its context, we also need to observe token within a sentence. For example, consider the following sentence:

**S3:** Pelaku **tabrak lari** yang menewaskan Siska **berhasil ditangkap** di jalan Monginsidi.  
(The *hit-and-run* actor who killed Siska was captured on the Monginsidi road.)

In S3, there are two possible phrases that can fill *what* slot, *tabrak lari* and *berhasil ditangkap*. Both phrases are preceded by a noun that can act as a subject. But, when we see the entire sentence, the phrase of *berhasil ditangkap* is *what* element and the phrase of *tabrak lari* describes the word of *pelaku* which is also the *who* element.

Since this 5W1H extraction is a sequence labelling task, we use bidirectional long short-term memory (BLSTM). BLSTM can incorporate past and future input information [17]. Network steps through the input sequence in both directions, backward and forward.

To induce sentence-level features, we use convolutional neural network, which has been successfully applied to event detection [7,8]. Output from BLSTM units and max pooling layer in convolutional neural network are combined with another additional features and then are fed into classifier. figure 1. illustrates the proposed network architecture.



**Figure 2.** Convolutional Neural Network. It illustrates the processing of token *menandatangani* from sentence “*Jokowi menandatangani surat perjanjian kerjasama.*”. We use two filter with window size in the set  $\{2,3,4\}$  to generate feature maps, and apply max pooling to each feature map.

### 3.1. Extracting Sentence Features with Convolutional Neural Network

Convolutional neural network can learn semantically meaningful representations of sentences. With max pooling layer, we can get the most useful information that represents the entire sentence. Figure 2. describes our convolutional neural network in details. This convolutional neural network is inspired by [7,8].

**3.1.1. Input.** In the CNN, an input must have fixed size. Therefore, we limit the context to a fixed window size by trimming longer sentence and padding shorter sentence with 0 when necessary. Before entering convolutional layer, each token is transformed into real-valued vector by using the following embedding:

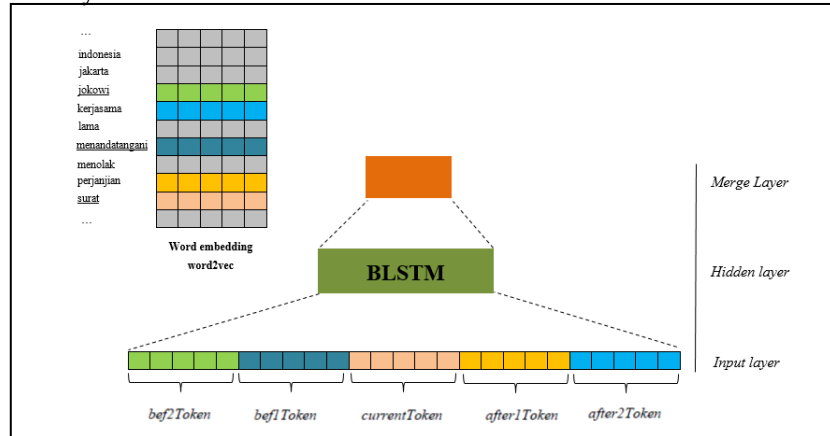
- **Word Embedding.** We use word2vec [16] embedding trained on 10.956 Indonesian news articles. Each token in a sentence is transformed into 50-dimensional word vector. If token was not found in word embedding, we randomly initialized the word embedding vectors using uniform distribution with the interval from -0.25 to 0.25.
- **Position Embedding.** To specify the token that will be predicted, we use position embedding. This embedding randomly initialized. Position embedding provides information about relative distance each token within a sentence with current token.

We concatenate real-valued vector from word embedding and position embedding into one 2-dimensional matrix. We then fed this input matrix into convolutional layer.

**3.1.2. Convolutional Layer.** In this layer, vector matrix will be convoluted by filter with some window size  $k$ . This filter is composed by weight vectors and bias, and slides over the input matrix. After

operating non-linear function, we obtained some valuable features with lower dimension in feature map.

**3.1.3. Max Pooling Layer.** The pooling units perform max operation to capture the most important features (max value) within each feature map. These features are concatenated with other features into single vector before they are fed into classifier.



**Figure 3.** Bidirectional LSTM for sequence labelling. It illustrates the processing of one sequence that consist of current token *surat*, bef1token *menandatangani*, bef2token *jokowi*, after1token *perjanjian* and after2token *kerjasama*.

### 3.2. Sequence Labelling with Bidirectional Long Short Term Memory.

Given a sequence  $x_{n-2}, x_{n-1}, x_n, x_{n+1}, x_{n+2}$  with label  $y$ . Bidirectional LSTM is used to model this sequence and to analyze the relationship among tokens in the sequence. Each token  $x$  is transformed into word vector by using word embedding. Output BLSTM will be combined with other features and then be fed into classifier. Figure 3. illustrates this BLSTM unit.

### 3.3. Additional Features

Other than all tokens and their relative position within a sentence (**SENT**) and lexical sequence (**LEX**), we use the following additional features:

- **LOCT:** Word's position within a sentence has a positive meaning for determining its role.

$$\text{LOCT} = 1 - \frac{\log t}{\log n} \quad (1)$$

Where  $n$  is the number of tokens in the sentence and  $t$  is token location within the sentence. For instance, if token  $x$  is at beginning of the sentence,  $t=1$ .

- **LOCS:** Sentence token location within an article. *Who*, *what*, *where*, and *when* elements often appear at the beginning of a news article. And *why* and *how* elements often appear at the middle or end of the article.

$$\text{LOCS} = 1 - \frac{\log s}{\log n} \quad (2)$$

Where  $n$  is the number of sentences in an article and  $s$  is sentence token location within the article.

## 4. Experiments

### 4.1. Dataset

We use 5W1H corpus of Indonesian news articles that have been collected by Wicaksana [14]. The corpus labeled manually by human annotator, consists of 484 news articles. We utilized test set with



145 news articles, and training set with the remaining 339 articles. Figure 4. shows an example of 5W1H information on Indonesian news article [13].

From 484 news articles, we built dataset of BIO-labeled token. After annotating, we obtained 123.586 instances. Table 1. shows the label distribution.

From the distribution labels, we can see that it is an imbalanced dataset. To overcome this problem, we apply random oversampling to minority classes in training set.

**Table 1.** Label Distribution Based on BIO Labelling Scheme

Label	Count	
	<i>Training Set</i>	<i>Testing Set</i>
<b>B_who</b>	340	145
<b>I_who</b>	1.586	713
<b>B-what</b>	340	145
<b>I_what</b>	3.924	1.418
<b>B_where</b>	403	241
<b>I_where</b>	921	509
<b>B_when</b>	324	142
<b>I_when</b>	1.300	624
<b>B_why</b>	302	120
<b>I_why</b>	3.863	1.583
<b>B_how</b>	324	122
<b>I_how</b>	5.164	2.060
<b>Other</b>	69.179	27.794

Rakyat Venezuela Mulai Pilih Pengganti Chavez [ <i>Venezuelans Vote to Choose Chavez Successor</i> ]	5W1H Information
REPUBLIKA.CO.ID, KARAKAS – Transisi kekuasaan akan dimulai di Venezuela. Tidak kurang dari <u>19 juta pemilih</u> , akan menentukan masa depan politik negara tersebut saat <u>Ahad (14/4)</u> waktu setempat. Beberapa hasil polling masih menempatkan, pejabat presiden sementara, Nicolas Maduro unggul dari pesaingnya, Henrique Capriles. Tapi, prediksi tersebut dikhawatirkan berbalik. Reuters mengatakan Maduro, menurut polling, unggul dua digit meninggalkan bekas Gubernur Negara Bagian Miranda tersebut. Namun, kesenjangan popularitas itu kian menyempit menjelang pemilihan. Dikatakan jarak popularitas hanya terpaut tujuh persen. Capriles mengatakan, aksi manipulasi politik Maduro telah memecah belah masyarakat Venezuela. Dirinya yakin akan mengambil peran presiden, and akan menjungkalkan rezim sosialisme di negara itu. Dalam kampanye, dia mengatakan, pesaingnya itu culas memanfaatkan mendiang Presiden Hugo Chavez untuk kepentingan pribadinya. Padahal, kata dia, mendiang Chavez adalah milik semua masyarakat Venezuela. Tempat pemungutan suara akan di buka Ahad (14/4) pukul 06:30 waktu setempat, atau Senin (15/4) waktu Indonesia. <u>TPS akan dibuka selama 10 jam, dengan voting secara elektronik. Pemenangnya akan dilantik 19 April mendatang, and akan menentukan pemerintahan sampai 2019 mendatang.</u>	<p><b>What:</b> menentukan masa depan politik negara tersebut [<i>determine political future of the country</i>]</p> <p><b>Who:</b> 19 juta pemilih [<i>19 million voters</i>]</p> <p><b>Where:</b> Venezuela</p> <p><b>When:</b> Ahad (14/4) [<i>Sunday, 14 April</i>]</p> <p><b>Why:</b> Pemenangnya akan dilantik 19 April mendatang, and akan menentukan pemerintahan sampai 2019 mendatang [<i>The winner is due to be sworn in on 19 April and serve until January 2019</i>]</p> <p><b>How:</b> TPS akan dibuka selama 10 jam, dengan voting secara elektronik. [<i>Polls will open for 10 hours, by electronic voting</i>]</p>

**Figure 4.** Example of 5W1H information on Indonesian news article [13]

#### 4.2. Hyperparameter

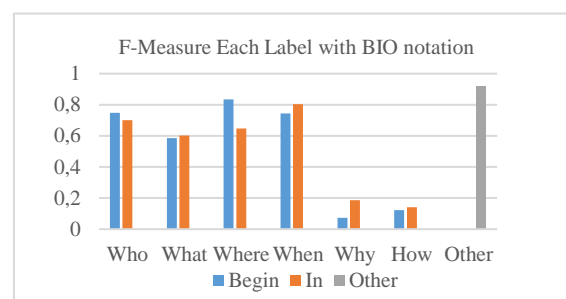
We train the network with batch size=64 and epoch=10. We used Adadelata [18] optimizer (lr=1.0, rho=0.95, decay=0.0, epsilon=1e-08). ReLU activation function was used on convolutional layer and fully connected layer. Because we convert the output variable into one-hot encoding representation, we apply sigmoid function on output layer.

For convolutional neural network, we use multiple window size 2, 3, 4 and utilized two features maps for each window size. For bidirectional LSTM, we set 250 units LSTM. In order to avoid

overfitting, we apply dropout [19]. Dropout is one of the regularization techniques that randomly turns off neurons during training phase. We set dropout rate 0.2 for input layer CNN, 0.3 for output layer CNN and fully connected layer, and 0.5 for BLSTM layer.

**Table 2.** Label Performance Based On Different Feature Set

Features	P	R	F
LEX (BLSTM)	0.698	0.641	0.667
SENT (CNN)	0.747	0.591	0.648
SENT + LEX (CNN+BLSTM)	0.783	0.831	0.8
+ additional features			
<b>SENT + LEX + LOCT + LOCS</b>	<b>0.798</b>	<b>0.828</b>	<b>0.808</b>



**Figure 5.** Performance based on BIO notation

#### 4.3. Performance

In our experiments, we use precision (P), recall (R), and F-measure (F) to evaluate extraction performance. In the first experiment, we test the performance when introducing different feature set. Table 2. shows the result.

The results reveal that the best performance obtained by using all features, SENT+LEX+LOCT+LOCS, with F-measure 0.808, Precision 0.798, and Recall 0.828. From the experimental results, lexical sequence BLSTM with F-measure 0.667 and then sentence semantic features resulted by CNN contributed to performance with F-measure 0.648. The combination of these two features gave F-measure 0.8.

Deep learning has the ability in feature learning that learns about the features of raw data through its algorithm so it does not depend complicated features which are defined manually by involve experts or

feature engineering [20]. This can be seen from the F-measure performance of the CNN + BLSTM model without the additional features is good enough in doing token classification. In addition of token location features (LOCT and LOCS) increase F-measure by 0.008 thus achieve the best performance with F-measure 0.808. The best performance was obtained without using features from NLP tools like POS tag and NE tag.

Figure 5. shows the model performance of the 13 labels. The highest value is shown by *when* element. Because the pattern of this element is most easily detected. When element, which denotes time, has a fixed pattern, consisting of a set of days and or numbers showing the date and time. These results are also shown in the previous researches [13,15], where *when* element is the element most easily detected and has the highest accuracy. *Why* and *how* elements have a low accuracy value.

The *why* and *how* elements, which provide an additional description of the event, are commonly contained in one sentence with the *who* and *what* elements, but more often found in separate sentences that are in the middle or end of the news. When the *why* and *how* elements are a separate sentence, the



*why* sentence can be detected through the use of certain words e.g. the word "because", "for", "caused by" which indicates the *why* element. But sometimes there are sentences that contain the word but not the *why* element. These sentences have no semantic relation to *what* element.

While the *how* sentence has no special characteristics, the characteristics of this sentence have similarities with the sentence of *who* and *what*. There should be information about the semantic relationship between *why* and *how* with *what* (the element that states the event) so that it can distinguish

**Table 3.** Extraction Performance of Naïve Bayes, C4.5, IBk, and CNN-BLSTM

Method	P	R	F
Naïve Bayes	0.702	0.526	0.595
C4.5	0.732	0.587	0.645
IBk	0.672	0.682	0.655
<b>CNN + BLSTM</b>	<b>0.798</b>	<b>0.828</b>	<b>0.808</b>

the sentence with a sentence labeled *other*. The absence of features that represent semantic relation among sentences for these two elements causes low accuracy.

#### 4.4. Performance Comparison

This experiment aims to compare the performance of CNN-BLSTM with other shallow methods i.e. IBk, C4.5, and Naïve Bayes on our dataset and features set. We cannot directly compare our results with the results in previous studies [13-15] because of differences in dataset. Table 3. shows the experimental results. We can conclude that the CNN-BLSTM method has the best performance with F-measure of 0.808. IBk, C4.5, and Naïve Bayes gave the average F-measure of 0.655, 0.645, and 0.595, respectively.

This demonstrates the reliability of deep learning in feature abstractions from high-dimensional data and the ability in generating feature functions from raw data thereby enhancing classification precision. Shallow learning is difficult to approach the functions of complex variables consequently resulting lower precision.

### 5. Conclusion and Future Works

5W1H extraction on Indonesian news article is a multiclass classification problem. By using BIO notation, each token in an article is classified into one of 13 classes, *B-who*, *I-who*, *B-what*, *I-what*, *B-when*, *I-when*, *B-where*, *I-where*, *B-why*, *I-why*, *B-how*, *I-how*, and *Other*. Label token depends on other token observations in the sequence and the information the entire sentence. We use CNN to extract information syntactic and semantic token within a sentence and bidirectional LSTM to learn the relations among tokens in the sequence. From six news elements, 5W1H, *when* elements are most easily detected, otherwise *why* and *how* elements are the most difficult ones.

Deep learning has the advantage of feature abstraction from high-dimensional data and the ability in generating feature functionality from the data thereby enhancing classification precision. While shallow learning cannot approach the function of complex variables. This is demonstrated by the results of tests comparing the performance of CNN-BLSTM with IBk, C4.5, and Naïve Bayes. The best performance was obtained by CNN-BLSTM with the F-measure 0.808. IBk, C4.5, and Naïve Bayes obtain the F-measure 0.655, 0.645, and 0.595, respectively.

For future works, we will improve model by adding number of training data. Deep learning algorithms perform better with more data. To sequential data modeling, we will employ development of LSTM unit, *Gated Recurrent Units* (GRU) and *Minimal Gate Unit* (MGU), to compare with the BLSTM in this task. Furthermore, we will explore more potential features and provide information that can represent semantic relation between *why* and *how* sentences with *what* elements and title/news headline so it can improve model performance in extracting *why* and *how* information.

## Acknowledgement

This research receives grant from Superior Scholarship (*Beasiswa Unggulan*) by Bureau of Planning and Foreign Cooperation, Ministry of Education and Culture RI.

## References

- [1] Jurafsky D and Martin J H 2008 *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2<sup>nd</sup> Edition (Upper Saddle River: NJ: Prentice-Hall)
- [2] Hinton G, Deng L, Yu D, George E D, Mohamed A-R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N and Kingsbury B 2012 *Deep Neural Network for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups* (IEEE Signal Processing Magazine vol. 29) pp 82 – 97
- [3] Yan Y, Chen M, Shyu M-L and Chen S-C 2015 *Deep Learning for Imbalanced Multimedia Data Classification, Int. Symp. on Multimedia* (Miami: IEEE)
- [4] Kalchbrenner N, Grefenstette E and Blunsom P 2014 A Convolutional Neural Network for Modelling Sentences *Proc. 52nd Annu. Mtg. Assoc. Comput. Linguistics (Baltimore)* pp 655–65
- [5] Kim Y 2014 Convolutional Neural Networks for Sentence Classification *Proc. Conf. EMNLP (Doha)* pp 1746–51
- [6] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K and Kuksa P 2011 Natural Language Processing (Almost) from Scratch *J. of Machine Learning Research* **12** pp 2493–2537
- [7] Chen Y, Xu L, Liu K, Zeng D and Zhao J 2015 Event Extraction via Dynamic Multi-Pooling Convolutional Neural Network *Proc. 53rd Annu. Mtg. Assoc. Comput. Linguistics (Beijing: Association for Computational Linguistics)* pp 67–176
- [8] Nguyen T H and Grishman R 2015 Event Detection and Domain Adaptation with Convolutional Neural Network *Proc. 53rd Annu. Mtg. Assoc. Comput. Linguistics (Beijing: Association for Computational Linguistics)* pp 365–71
- [9] Sundermeyer M, Alkhoul T, Wuebker J and Ney H 2014 Translation Modeling with Bidirectional Recurrent Neural Network *Proc. Conf. EMNLP (Doha, Qatar)* pp 14–25
- [10] Wang P, Qian Y, Soong F K, He L, Zhao H 2015 *Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network* (Cornell University Library) Preprint cs/1510.06168
- [11] Sundermeyer M, Ney H and Schluter R 2015 From Feedforward to recurrent lstm neural networks for language modelling *J. IEEE/ACM Trans. Audio Speech Lang. Process.* **23** issue 3 pp 517–29
- [12] Sundermeyer M, Schluter R and Ney H 2012 *LSTM Neural Networks for Language Modeling, Proc. ISCA Interspeech* (Portland) pp 194–7
- [13] Khodra M L 2015 Event Extraction on Indonesian News Article Using Multiclass Categorization *ICAICTA (Chonburi: IEEE)*
- [14] Wicaksana A P 2015 *Ekstraksi Informasi 5W1H Pada Berita Bahasa Indonesia* (Bandung: ITB)
- [15] Ilyas R and Khodra M L 2015 Ekstraksi Informasi 5W1H pada Berita Online Bahasa Indonesia *J. Cybermatika* **3** JC03(2015)006
- [16] Mikolov T, Chen K, Corrado G and Dean J 2013 Distributed Representations of Words and Phrases and their Compositionality *Proc. Int. Conf. Neural Inf. Process. Syst. (Lake Tahoe, Nevada)* pp 3111–9

- [17] Graves A, Fernández S and Schmidhuber J 2005 Bidirectional LSTM networks for improved phoneme classification and recognition *Proc. 15<sup>th</sup> Int. Conf. Artificial Neural Networks: Formal Models Appl.* vol II (Warsaw, Poland :Springer Berlin Heidelberg) pp 799-804
- [18] Zeiler M D 2012 *Adadelata: An Adaptive Learning Rate Method* (Cornell University Library) *Preprint* cs/1212.5701
- [19] Srivastava N, Hinton, Geoffrey E, Krizhevsky, Alex, Sutskever, Ilya and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. of Machine Learning Research* **15** issue 1 pp 1929–1958
- [20] Bengio Y 2012 *Deep Learning of Representations for Unsupervised and Transfer Learning*, *Proc. of the 2011 Int. Conf. on Unsupervised and Transfer Learning workshop* vol 27 (Washington: JMLR) pp 17-37