Literature Review: Information Extraction using Named-Entity Recognition with Machine Learning Approach

R Fenny Syafariani[a], Rio Yunanto[b]

[ab]Universitas Komputer Indonesia, Jl. Dipatiukur No. 112-116, Bandung, Indonesia
[a]r.fenny.syafariani@email.unikom.ac.id, [b]rio.yunanto@email.unikom.ac.id

Abstract

The purpose of this study is to help researchers identify and map machine learning algorithms from the results of previous studies with the theme of recognizing named-entities. This study's research method examines works of literature on the topic of introducing named-entities with the machine learning approach. The literature ranged from the year 2018 to 2020 and was collected through the use of Google Scholar. In this study, one of the critical research questions to be answered is whether machine learning algorithms have been used in named-entity recognition research. The introduction of named-entities is able to use three approaches: 1) machine learning, 2) deep learning, and 3) a combination of both. From the result, it was discovered that the combination of Conditional Random Field (CRF) machine learning and Bidirectional Long Short-Term Memory (Bi-LSTM) deep learning were used in 4 out of 7 analyzed works of literature.

Keywords: NER;information;extraction;named-entity;review;

1. Introduction

Entity extraction process is widely known to be one of the important stages in information extraction. As one of the methods, named-entity recognition can automatically extract entities in a particular text and determine its category. It includes extracting object name, object, person, or company name (Wibisono & Khodra, 2018). As an example, from the sentence "Flood and landslide in Nganjuk, 23 people reported missing", the recognition process will result in a named-entity (often referred to as a mention) with "Nganjuk" as the type of location as well as "Flood" and "landslide" as the type of event. It shows that the named-entity recognition process is able to automatically recognize entities in a sentence or text and is able to categorize the entity according to the type referred to in the text.

One of the ways that named-entity recognition can be done is through the formulation of a certain word or phrase patterns. For example, the typical word pattern of the phrase “come from…” or “go away from…” would be followed by location-type entity words. Various combinations of word patterns can be taught in machine learning using training data to build knowledge on the algorithms used. Therefore, it further supports the fact that the introduction of machine learning-based named-entity recognition will be able to detect named entities automatically (Giarsyani, 2020).

We have conducted a literature study to determine a suitable machine learning algorithm that could open up new research areas opportunities. Through literature study, we have created research questions as a guide in the research process which includes: 1) What objects or datasets have been used in the research on recognizing named-entity?, 2) What machine learning algorithms have been used in named-entity recognition research?, and 3) What are the results of applying machine learning algorithms in the research on named-entity recognition?

2. Method

A literature study was conducted to identify and map the results of previous studies related to certain literature themes. In addition, a good literature study will produce a map of knowledge about a research topic that can guide researchers to dig deeper into areas that are not yet mature (Fisch & Block, 2018). The literature data in this study were collected through the use of Google Scholar with the keyword "Named Entity Recognition". The literature with the topic of introducing named-entity was then selected according to several factors, namely: 1) The approach used only focuses on the machine learning approach, and 2) The publication year of the literature obtained should be from the year 2018 to 2020. The results of the gradual selection resulted in seven pieces of literature which will be used as materials for comparison.

3. Results and Discussion

The process of analyzing and extracting large amounts of unstructured text or documents using Artificial Intelligence algorithms is often referred to as text mining. One part of text mining is the process of recognizing named-entities that can be used in various fields such as economy, health, social, politics, or culture. Based on the seven pieces of literature analyzed in this study, six pieces of literature apply the introduction of the main entity in the health sector, especially in the field of biomedicine and medicine. On the other hand, Wintaka's research used data taken from Twitter social media to identify the entity's name, location name, and organization name (Wintaka et al., 2019). The pieces of literature used in this research are shown in Table 1.

The health sector, especially the pharmaceutical industry, requires research on the introduction of named-entities, especially the medicine entities. The influence of a particular medicine with other medicines is closely monitored by the pharmaceutical industry in order to maintain patient safety from side effects caused by drug interactions (Chukwuocha et al., 2018). The biomedical field also has a very large corpus and requires information extraction to reduce the ambiguity due to several different entities that have the same acronym. Furthermore, several biomedical entities have inconsistent use of prefixes and suffixes (Cho et al., 2020).

Table 1. Literature Review Data based on Dataset

No	Ref	Year	Object / Dataset	Machine learning
1	(Chukwuocha et al., 2018)	2018	Medicine names / PubMed dataset	Conditional Random Field (CRF), and Naive Bayes (NB)
2	(Phan et al., 2019)	2019	Biomedical texts / BioNLP 2004 Challenge dataset	Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN)
3	(Casillas et al., 2019)	2019	Medical Online Corpus (GEN-MED) IXAMed Spanish EHR Corpus (EHR)	Bidirectional Long Short-Term Memory (Bi-LSTM), and Conditional Random Field (CRF)
4	(Suárez-Paniagua et al., 2019)	2019	eHealth-KD dataset	Bidirectional Long Short-Term Memory (Bi-LSTM), and Conditional Random Field (CRF)
5	(Wintaka et al., 2019)	2019	600 manually-labeled tweets in Bahasa Indonesia from Twitter social media	Bidirectional Long Short-Term Memory (Bi-LSTM), and Support Vector Machine (SVM)
6	(Gligic et al., 2019)	2019	Informatics for Integrating Biology & the Bedside – i2b2 dataset (2007-2012)	Forwards Neural Network (FFN), and Recurrent Neural Network (RNN), and Bidirectional Long Short-Term Memory (Bi-LSTM)

7	(Cho et al., 2020)	2020	JNLPBA (Biomedical) dataset NCBI (National Center for Biotechnology Information) dataset	Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (Bi-LSTM)

From Table 1, the popular machine learning algorithm used is the Conditional Random Field (CRF), while the popular deep learning algorithm is Bidirectional Long Short-Term Memory (Bi-LSTM). CRF is one of the various algorithms that are known to be great in building predictive models. CRF with its probabilistic model can be used for pattern recognition because it can consider word order labels that form sentences to identify entities from a text (Casillas et al., 2019). The LSTM algorithm is a development of the Recurrent Neural Network (RNN) algorithm through the generation of a memory cell that functions as a container for information for a long period. As for the Bi-LSTM algorithm, it has two layers that can move forward and backward. Bi-LSTM algorithm is generally used to handle sequential data to improve prediction accuracy (Cho et al., 2020).

The combination of Bi-LSTM and CRF approach is shown in Figure 1 (a), it shows two modules that compose a two-stage information extraction system. The input for the first Bi-LSTM layer is word embedding, in which the obtained output from the first layer is combined with word embeddings and sense-disambiguous embeddings in the second layer. Additionally, CRF was used in the final stage to get the most appropriate label for each token (Suárez-Paniagua et al., 2019). The concept of Medical Entity Recognition (MER) as shown in Figure 1 (b), relates to natural language processing applied to the clinical domain. The combination of Bi-LSTM with CRF serves to adapt the sequential tagger and to make it tolerant of high lexical variability and a limited number of corpus (Casillas et al., 2019).

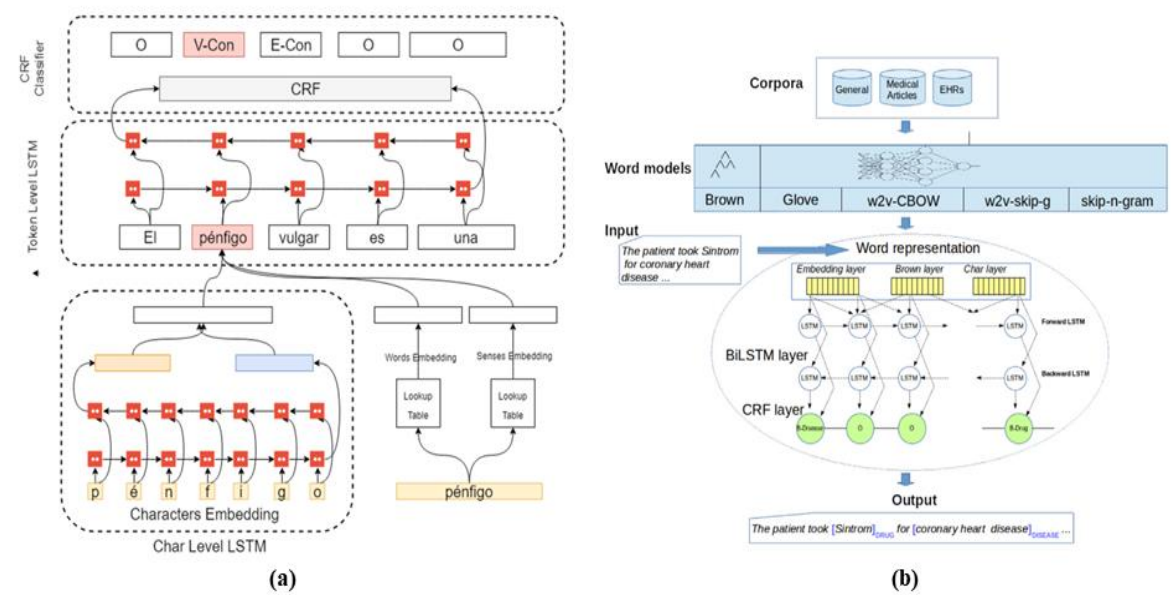

Fig. 1. (a) Two-stage deep learning approach (Casillas et al., 2019); (b) Neural architecture based on Bi-LSTM and CRF (Suárez-Paniagua et al., 2019);

A similar NER model but applied to different natural languages is still a frequent problem and it is still necessary to embed different trained words for each different natural language. As the first step, choosing the right algorithm and continuing to choose the corpus domain and genre is crucial to the success of the research.

Behind the various advantages of Bi-LSTM, there is still a problem where the complex Bi-LSTM algorithm architecture becomes one of the high computational burdens when applied to large-scale cases.

4. Conclusions

Literature studies of the previous researches obtained seven pieces of literature published from 2018 to 2020 with the research theme of the introduction of named-entity that can use 3 approaches, namely: 1) Machine learning, 2) Deep learning, and 3) A combination of machine learning with deep learning. The combination of machine learning and deep learning was used in 4 studies from 7 analyzed pieces of literature, namely the combination of Bidirectional Long Short-Term Memory (Bi-LSTM) deep learning with Conditional Random Field (CRF) machine learning. CRF with its probabilistic model can be used for pattern recognition because it is able to consider word order labels while the two layers of Bi-LSTM can handle sequential data to improve prediction accuracy by moving forward and backward. Currently, we are interested in exploring the topic of fake news detection but have not conducted any specific experiments related to NER. After conducting this literature study, we plan to explore NER using a fake news dataset that researchers have not been done before.

Acknowledgements

The authors wish to thank the Faculty of Engineering and Computer Science Universitas Komputer Indonesia for technical support. The Research presented in this paper has been done in the Laboratory of Accounting Information Systems, Universitas Komputer Indonesia.

References

Casillas, A., Ezeiza, N., Goenaga, I., Pérez, A., & Soto, X. (2019). Measuring the effect of different types of unsupervised word representations on Medical Named Entity Recognition. *International Journal of Medical Informatics*, *129*, 100–106. https://doi.org/10.1016/j.ijmedinf.2019.05.022

Cho, M., Ha, J., Park, C., & Park, S. (2020). Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics*, *103*, 103381. https://doi.org/10.1016/j.jbi.2020.103381

Chukwuocha, C., Mathu, T., & Raimond, K. (2018). Design of an interactive biomedical text mining framework to recognize real-time drug entities using machine learning algorithms. *Procedia Computer Science*, *143*, 181–188. https://doi.org/10.1016/j.procs.2018.10.374

Fisch, C., & Block, J. (2018). Six tips for your (systematic) literature review in business and management research. *Management Review Quarterly*, *68*(2), 103–106. https://doi.org/10.1007/s11301-018-0142-x

Giarsyani, N. (2020). Komparasi Algoritma Machine Learning dan Deep Learning untuk Named Entity Recognition : Studi Kasus Data Kebencanaan. *Indonesian Journal of Applied Informatics*, *4*(2), 138. https://doi.org/10.20961/ijai.v4i2.41317

Gligic, L., Kormilitzin, A., Goldberg, P., & Nevado-Holgado, A. (2019). Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. *ArXiv*, *121*, 132–139.

Phan, R., Luu, T. M., Davey, R., & Chetty, G. (2019). Biomedical named entity recognition based on hybrid multistage cnn-rnn learner. *Proceedings - International Conference on Machine Learning and Data Engineering, ICMLDE 2018*, 136–141. https://doi.org/10.1109/iCMLDE.2018.00032

Suárez-Paniagua, V., Rivera Zavala, R. M., Segura-Bedmar, I., & Martínez, P. (2019). A two-stage deep learning approach for extracting entities and relationships from medical texts. *Journal of Biomedical*

Informatics, *99*, 103285. https://doi.org/10.1016/j.jbi.2019.103285
Wibisono, Y., & Khodra, M. L. (2018). Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin. *INA-Rxiv*. https://doi.org/10.31227/osf.io/vud2p
Wintaka, D. C., Bijaksana, M. A., & Asror, I. (2019). Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF. *Procedia Computer Science*, *157*, 221–228. https://doi.org/10.1016/j.procs.2019.08.161