Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 425–432

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Computer Science and Computational Intelligence 2018

Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs

William Gunawan, Derwin Suhartono*, Fredy Purnomo, Andrew Ongko

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
Email: william.gunawan001@binus.ac.id, dsuhartono@binus.edu, fpurnomo@binus.edu, andrew.ongko@binusian.org

Abstract

In this paper, we describe the implementation of Named-Entity Recognition (NER) for Indonesian Language by using various deep learning approaches, yet mainly focused on hybrid bidirectional LSTM (BLSTM) and convolutional neural network (CNN) architecture. There are already several developed NERs dedicated to specific languages such as English, Vietnamese, German, Hindi and many others. However, our research focuses on Indonesian language. Our Indonesian NER is managed to extract the information from articles into 4 different classes; they are Person, Organization, Location, and Event. We provide comprehensive comparison among all experiments by using deep learning approaches. Some discussions related to the results are presented at the end of this paper. Through several conducted experiments, Indonesian NER has successfully achieved a good performance.

© 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Selection and peer-review under responsibility of the 3rd International Conference on Computer Science and Computational Intelligence 2018.

Keywords: Named-Entity Recognition; deep learning; natural language processing

1. Main text

Named-Entity Recognition (NER) is one subtask of Natural Language Processing research which is also included in Artificial Intelligence (AI) field [1]. NER is very useful to extract information from text by identifying and recognizing the entity of such person, organization, location, etc. [2]. NER is valuable as the ingredients of several NLP tasks. It has

* Corresponding author. Tel.: +6221-534-5830; fax: +6221-530-0244.
E-mail address: dsuhartono@binus.edu

1877-0509 © 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Selection and peer-review under responsibility of the 3rd International Conference on Computer Science and Computational Intelligence 2018.
10.1016/j.procs.2018.08.193

been successfully implemented in many areas, such as translation engine, search engine, automatic document indexing, automated question answering system, information retrieval, and so on [3].

NER in English shows an outstanding result due to the abundance of available data in conducting the research [4]. Yet, the experiment of NER in Indonesian language is not that good. This may be caused by the nature of Indonesian language. It varies in many ways due to the morphology history. Moreover, if we are going to use deep learning, lack of available dataset will become a major issue of the research.

Many approaches have been conducted to obtain significant results by using machine learning algorithm as well as rule-based. However, machine learning approach still has limitations for several languages [1]. For example, in term of its usage, Indonesian language has unique orthographic, morphological, and local contextual characteristics in both formal and informal [5]. By looking to aforementioned facts, it is a challenge and not an easy effort to manage NER for particular languages.

In general, NER is used to extract and gain information from documents by recognizing the entities of each word. Researchers attempt to develop Indonesian NER by using supervised learning [6], semi-supervised learning [7], rule-based [8], hybrid approaches [4], and using word vectors [9]. Furthermore, Indonesian NER plays an important role in health sector as well, such as recognizing part of medical entities [6] [9].

This research is highly inspired by some success story of previous approaches. Some of them are end-to-end sequence labeling [10], GRAM-CNN [11], deep active learning [12], NER for Nepali language [13], end-to-end RNN [14], residual LSTM [15], and highly inspired by BLSTM-CNNs approach [16]. To the best of our knowledge, this research is the first Indonesian NER that utilizes deep learning. By involving deep learning approaches combined with n-gram by CNN [11], we can reduce several steps to manually generate some features.

2. Related Works

There has been a lot of research on NER. The study was initialized in 1996 [3]. It was conducted by using a rule-based approach. As time goes by, NER research grows a lot and it involves many factors to consider [3]. Some example of NER research which involve statistical approaches are hidden Markov model [1] [17], maximum entropy Markov model [18], semi-Markov model [19], etc.

The first Indonesian NER study began in 2005 [8]. In recent years, the hype of NER research for Indonesian language has arrived. This research was conducted by using data from Indonesian DBPedia as labelled data and news article from Indonesian news sites (*kompas.com, cnnindonesia.com, tempo.co, merdeka.com* and *viva.co.id*) as unlabelled data with semi-supervised learning approaches which achieved 76.5% as the best accuracy score [7]. The other research used corpus from Indonesian news articles (*kompas.com* and *republika.com*) using rule-based approaches and achieved 67.37% as the best accuracy score [8]. Other research used corpus from 457 news articles that belongs to 3 categories from Detik (*detik.com*), Kompas (*kompas.com*) and Media Indonesia (*mediaindonesia.com*). It obtained 52.8% as the best accuracy score using ensemble supervised learning approach [4].

Deep learning model in NER is very popular in other countries, yet it does not happen with research tasks in Indonesia. Several languages that have been working on NER by using deep learning are Vietnamese [14], English [12], Chinese [20], French [21], Korean [22], and Italian [23]. They use different architectures along with the various nature of each language.

3. Neural Network Architecture

In this section, we describe the layers of our neural network architecture. The neural layers in our neural network are described one-by-one from bottom to top as seen in figure 1.

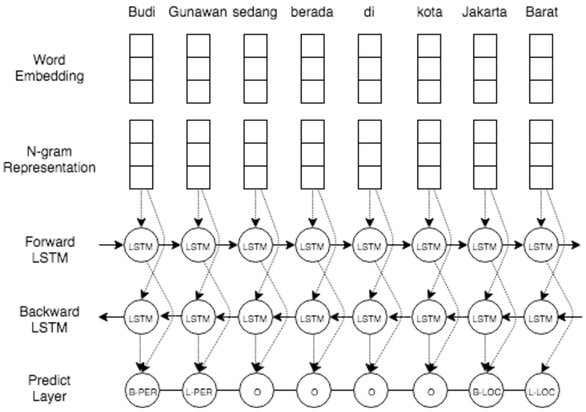


Fig. 1. Neural network architecture using hybrid BSLTM-CNNs.

As described in figure 1, our architecture starts from the word embedding layer. On the second step, N-gram representation is constructed and processed by using CNN layer. BLSTM means that the LSTM model is built in two directions: forward and backward. This layer is attached after CNN layer. Finally, fully-connected layer is put to predict the result.

3.1 Word Vector Representation

One example of NER usage in medical texts is to obtain important medical information, like diseases, symptoms, and drugs. This research compares 2 kinds of word representations; they are cluster-based word representation and distributed representation [9]. Word vectors are said to be successful in various NLP tasks, such as NER, POS tagging and dependency parser. Each word is assigned a dense low-dimensional real-valued vector, also called an embedding [24].

In the first layer, we used word embeddings. The word embeddings was trained using vector space models with Skip-gram approach. Distributed representations of words in a vector space help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words [25]. We run the experiment on 2 (two) different embeddings dimension. They are 100-dimensions and 200-dimensions embeddings.

3.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) is an effective approach to extract morphology information from sentences or words [14]. CNN plays an important role for many tasks of deep model architecture. It works very well to extract information from each sequences of words. In this research, CNN layers work to extract morphological information from a given sentence. CNN layer was applied after dropout layer [26], without any additional of handcrafted features.

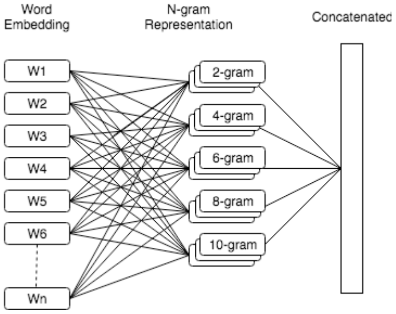


Fig. 2. Generated multi-sequence from N-gram.

By using CNN, n-gram is treated as n-word from a sentence. In this case, we generate 5 different n-gram language model; they are 2-gram, 4-gram, 6-gram, 8-gram and 10-gram. N-gram is used to build a multi-sequence of the words given from a sentence. Collection of generated n-gram which is called as multi-sequence is described in figure 2. Vectors from multi-sequence of the words are concatenated forming a new vector. This vector is used as the input in BLSTM block.

3.3 Long Short-Term Memory

LSTM (Long Short-Term Memory) was invented in 1997 [27]. LSTM is frequently used as a model to solve problems in machine learning [28]. It is designed to handle long-term dependency problem that vanilla Recurrent Neural Network (RNN) is not able to accommodate [29]. Our motivation in using LSTM is to extract the information from sequential data. Scheme of LSTM unit is described in figure 3.

The mathematical formulas in LSTM at time (t) are:

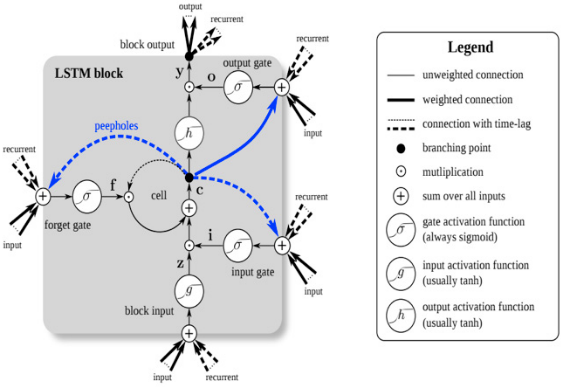


Fig. 3. Scheme of LSTM unit [10].

$$\begin{aligned}
i_t &= \sigma(W_i \mathrm{h}_{t-1} + U_i \mathrm{x}_t + b_i) \\
\mathrm{f}_t &= \sigma(W_f \mathrm{h}_{t-1} + U_f \mathrm{x}_t + b_f) \\
\tilde{\mathrm{c}}_\mathrm{t} &= \tanh(\mathrm{W_c} \mathrm{h}_t + U_f \mathrm{x}_t + b_c) \\
\mathrm{c}_\mathrm{t} &= \sigma(\mathrm{f}_t \odot \mathrm{c}_{t-1} + \mathrm{i}_t \odot \tilde{\mathrm{c}}_t) \\
\mathrm{o}_\mathrm{t} &= \sigma(W_o \mathrm{h}_{t-1} + U_o \mathrm{x}_t + b_o) \\
\mathrm{h}_\mathrm{t} &= \mathrm{o}_t \odot \tanh(\mathrm{c}_t)
\end{aligned}$$

3.3.1 Bidirectional-LSTM

For many sequences labeling tasks, it is beneficial to have access to both past (left) and future (right) [10]. The basic idea is to present each sequence forwards and backwards to two separated hidden states to capture past and future information, respectively. We acquired the best model with settings of 100 as the number of the LSTM units and 0.5 rates for both dropout and recurrent dropout.

3.4 Fully Connected Layers

Fully connected layers are the last part of layers in our model. Some deep learning architectures use dense, Conditional Random Fields (CRF), or other techniques for the final layer. In our case, we use dense layer with softmax as the activation function. Softmax function is well-known as a function that has a good performance to handle multi-class classification problem [30]. The formula of softmax function is described as follows:

$$P(\mathcal{Y}|x) = \frac{exp\left(w_{\mathcal{Y}}^{t} x\right)}{\sum_{\mathcal{Y}'} exp\left(w_{\mathcal{Y}'}^{t} x\right)}$$

The formula above is equivalent to apply a linear operation. Training amounts to adapting the vectors $w_{\mathcal{Y}}$ by maximizing the average conditional log-probability $\frac{1}{N}\sum_{\alpha} \log p\,(\mathcal{Y}^{\alpha}|x^{\alpha})$ for a set $\{(x^{\alpha}, \mathcal{Y}^{\alpha})\}_{\alpha=1}^{N}$ of training cases [31].

4. Named-Entity Classes

The structure of the label scheme is "BILOU", which stands for **B**eginning, the **I**nside, and **L**ast token of multi-token and it stands for **O**utside and **U**nit for the single tokens. This label scheme has been introduced in previous research. Some of NER research usually use BIO scheme or BILOU scheme. After conducting some experiments to compare the performance of both schemes, it was concluded that BILOU scheme significantly outperforms the widely adopted BIO [32].

In this research, we focused on some important entities, such as name of the Person, Location, Organization, and Event. For example, given the sentence:

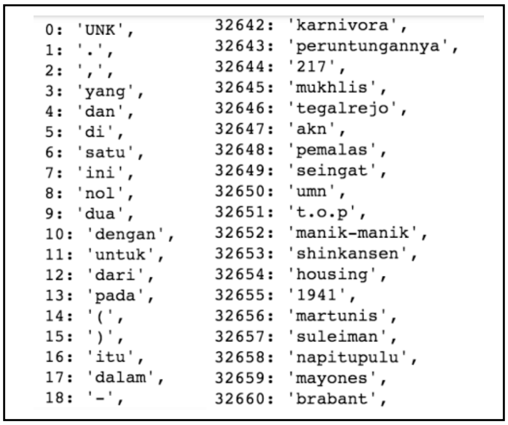
`"Joko Widodo sedang berada di depan Istana Presiden untuk bertemu dengan Komisi Pemberantasan Korupsi (KPK)."`
(literally means, "`Joko Widodo is in front of the Presidential Palace to meet the Corruption Eradication Commission`")

`Joko` will be predicted as "B-PERSON"; `Widodo` as "L-PERSON", `Istana` as "B-LOC", `Presiden` "L-LOC", and the rest will be the **O**utside entities that are labelled with "O". We have 4 core classes namely person, location, organization, and event. Each class has 4 subclasses which starts with "B", "I", "L", or "U". Thus, the total class is 17 classes, one of them is "O" (The Outside of entity class).

5. Corpus

The corpus consists of 4,139 sentences. It comes from news articles and some other articles about history of Indonesia taken from Wikipedia. The amount of words is 81,173. It consists of 10,889 words of entity classes and 70,284 words of non-entity classes ("O"). We use 4-fold cross validation to measure the performance of labelling the entities.

We use 700,000 Indonesian news articles and Indonesian Wikipedia dump file to train the word embeddings. Based on the number of occurrences, we choose the first 1,000,000 words (including punctuation) to be the vocabulary of our experiments. The example of vocabulary content is on figure 4. While the data that is not in our vocabulary will be predicted as non-entity class (labelled as **O**utside).



```
0: 'UNK',          32642: 'karnivora',
1: '.',            32643: 'peruntungannya',
2: ',',            32644: '217',
3: 'yang',         32645: 'mukhlis',
4: 'dan',          32646: 'tegalrejo',
5: 'di',           32647: 'akn',
6: 'satu',         32648: 'pemalas',
7: 'ini',          32649: 'seingat',
8: 'nol',          32650: 'umn',
9: 'dua',          32651: 't.o.p',
10: 'dengan',      32652: 'manik-manik',
11: 'untuk',       32653: 'shinkansen',
12: 'dari',        32654: 'housing',
13: 'pada',        32655: '1941',
14: '(',           32656: 'martunis',
15: ')',           32657: 'suleiman',
16: 'itu',         32658: 'napitupulu',
17: 'dalam',       32659: 'mayones',
18: '-',           32660: 'brabant',
```

Fig. 4. Sample of vocabulary from choosing the first 1,000,000 words

6. Results and Evaluation

As depicting in table 1 below, the highest performance is achieved by token "PERSON" while the lowest one is achieved by token "EVENT". Single token ('U') got better F1-Score than multiple token ('B', 'I', 'L'). This happened due to the condition that single token can be easily recognized by seeing its contexts. Vice versa, It is more complicated for recognizing multiple token. The understanding is not only required to the contexts of the sentence, but also the words inside the multiple token as well.

Table 1. Performance measurement to all labels of our model

Label	Precision	Recall	F1-Score	Frequency
U-PERSON	91%	82%	86%	1489
B-PERSON	86%	83%	84%	971
I-PERSON	74%	80%	77%	262
L-PERSON	89%	88%	89%	964
U-ORG	80%	81%	80%	991
B-ORG	74%	79%	76%	753
I-ORG	73%	78%	76%	539
L-ORG	72%	80%	76%	750
U-LOC	82%	81%	81%	1520
B-LOC	76%	80%	78%	904
I-LOC	67%	68%	67%	391
L-LOC	78%	83%	81%	900
U-EVENT	43%	19%	26%	46
B-EVENT	62%	58%	60%	142
I-EVENT	65%	42%	51%	124
L-EVENT	60%	55%	57%	143
O	98%	98%	98%	70284

The label ‘O’ always has a good performance because the data consists of more words compared to other labels. Each model in table 2 contains several entities with better classification values compared to BLSTM-CNNs. Yet, some models failed to recognize U-EVENT entities. Thus, BLSTM-CNNs can recognize all entities with small amount of data, even though the prediction score is still not adequate. After all experiments are conducted, we find out the best F-Score is 79.43% as described in table 2.

Table 2. Performance of Deep Learning Model

Model	Precision	Recall	F-Score
BLSTM	79.12%	74.79%	76.47%
BLSTM-CNNs	79.72%	79.61%	79.43%
BLSTM-LSTM	78.68%	76.86%	77.47%
BLSTM-CNNs-LSTM	76.63%	75.54%	75.83%

From the whole experiment, the performance of several combination of layers are shown in Table 2. We achieved the highest score by using BLSTM-CNNs model. More complicated models do not guarantee that it will get the best performance. Performance will be better if the model got tailored carefully to the data.

Sequence and complexity of a model impacts carry effects to the results. By observing the model comparison in table 2, BLSTM-CNNs achieved the best score among all other models. A single BLSTM model cannot surpass BSLTM-CNNs because the hybrid BLSTM-CNNs extracted more information compared to single BLSTM model. LSTM model that is combined in BSLTM-LSTM model did not have much effect to the model because computation in LSTM has been included in BSLTM layer. A more complex model, BLSTM-CNNs-LSTM got the smallest score among all other models. This happened due to the small amount of dataset which cannot accommodate a complex model to learn.

7. Conclusion

The experiment results reported that deep learning can achieve good performance even though the F1-score is quite low for some labels. It is possibly happened due to small number of dataset used in this research. BLSTMS-CNNs is considered as the best model for named-entity recognition task in Indonesian language.

For future study, we plan to collect and build more dataset. By having huge size of dataset, we should get better performance. It is reflected by the result of "PERSON" entity and "EVENT" entity. We also plan to modify the deep learning architecture such that the performance will get better.

8. Acknowledgment

Thanks to PT. Kurio for supports the data of this research.

References

1. Morwal S, Jahan N, Chopra D. Named Entity Recognition using Hidden Markov Model (HMM). In International Journal on Natural Language Computing (IJNLC); 2012: Semantic Scholar.
2. Alfred R, Leong LC, On CK, Anthony P. Malay Named Entity Recognition Based on Rule-Based Approach. In International Journal of Machine Learning and Computing; 2014: Semantic Scholar.
3. Sekine S. Named Entity: History and Future. 2003.
4. Wibawa AS, Purwarianti A. Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning. In Procedia Computer Science; 2016: Elsevier. p. 221-228.
5. Cohn AC, Ravindranath M. Local Languages in Indonesia: Language Maintenance of Language Shift? Masyarakat Linguistik Indonesia. 2014 August; 34(2): p. 131-148.
6. Suwarningsih W, Supriana I, Purwarianti A. ImNER Indonesian Medical Named Entity Recognition. In 2nd International Conference on Technology, Informatics, Management, Engineering & Environment; 2014; Bandung. p. 184-188.
7. Aryoyudanta B, Adji TB, Hidayah I. Semi-supervised learning approach for Indonesian Named Entity Recognition (NER) using co-training algorithm. In International Seminar on Intelligent Technology and Its Applications; 2016; Lombok: IEEE.
8. Budi I, Bressan S, Wahyudi G, Hasibuan ZA, Nazief BAA. Named entity recognition for the Indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach. In Discovery Science; 2005; Heidelberg: Springer. p. 57-69.
9. Rahman A. Medical Named Entity Recognition for Indonesian Language Using Word Representations. In IOP Conference Series: Materials Science and Engineering; 2018: IOP.
10. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. arXiv:1603.01354. 2016 Mar.
11. Zhu Q, Li X, Conesa A, Pereira C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics. 2018 May 1; 34(9): p. 1547-1554.
12. Shen Y, Yun H, Lipton ZC, Kronrod Y, Anandkumar A. Deep Active Learning for Named Entity Recognition. In 2nd Workshop on Representation Learning for NLP; 2017 july; Vancouver: Association for Computational Linguistics. p. 252-256.
13. Dev A, Prukayastha BS. Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach. International Journal of Computer Applications. 2013; 84(9).
14. Pham TH, Le-Hong P. End-to-end Recurrent Neural Network Models for Vietnamese Named Entity Recognition: Word-level vs. Character-level. arXiv:1705.04044. 2017 May.

15. Tran Q, MacKinlay A, Jimeno Yepes A. Named Entity Recognition with stack residual LSTM and trainable bias decoding. eprint arXiv:1706.07598. 2017 June.

16. Chiu JPC, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs. arXiv:1511.08308. 2015 November.

17. Setiyoaji A, Muflikhah L, Fauzi MA. Named Entity Recognition Menggunakan Hidden Markov Model dan Algoritma Viterbi pada Teks Tanaman Obat. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2017 December; 1(12): p. 1858-1864.

18. McCallum A, Freitag D, Pereira FCN. Maximum Entropy Markov Models for Information Extraction and Segmentation. In International Conference on Machine Learning; 2000; San Francisco: ACM. p. 591-598.

19. Sarawagi S, Cohen WW, Kou Z. Semi-Markov Models for Named Entity Recognition. In.

20. Peng N, Dredze M. Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. arXiv:1603.00786. 2016 March.

21. Gridach M, Mulki H, Haddad H. FNER-BGRU-CRF at CAp 2017 NER challenge: Bidirectional GRU-CRF for French Named Entity Recognition in Tweets. In CAp; 2017; Grenoble.

22. Cheon MA, Kim CH, Park HM, Kim JH. Evaluating and applying deep learning-based multilingual named entity recognition. In Journal of the Korean Society of Marine Engineering; 2018. p. 106-113.

23. Bonadiman D, Severyn A, Moschitti A. Deep Neural Networks for Named Entity Recognition in Italian. In CLiC it; 2015. p. 51-55.

24. Hsieh JT, Li C, Liu W. Effective Word Representation for Named Entity Recognition. In ; 2017: Semantic Scholar.

25. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. 2013 October.

26. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. The Journal of Machine Learning Research. 2014 January; 15(1).

27. Hochreiter S, Schmidhuber J. Long Short-Term Memory. In Neural Computation; 1997; Cambridge: MIT Press. p. 1735-1780.

28. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: A Search Space Odyssey. arXiv:1503.04069. 2015 March.

29. Suhartono D, Gema AP, Winton S, David T, Fanany MI, Arymurthy AM. Attention-Based Argumentation Mining. Special Issue on: MIWAI 2017 Computational Intelligence and Deep Learning for Computer Vision, International Journal of Computational Vision and Robotics. 2017.

30. Brébisson Ad, Vincent P. An Exploration of Softmax Alternatives Belonging to the Spherical Loss Family. arXiv:1511.05042. 2015 November.

31. Memisevic R, Zach C, Pollefeys M, Hinton GE. Gated Softmax Classification. In Neural Information Processing Systems (NIPS); 2010: Curran Associates, Inc. p. 1603-1611.

32. Ratinov L, Roth D. Design Challenges and Misconceptions in Named Entity Recognition. In CoNLL '09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning; 2009; Boulder: Association for Computational Linguistics. p. 147-155.

33. Jiampojamarn S, Cercone N, Kešelj V. Biological Named Entity Recognition Using n-grams and Classification Methods. In Pacific Association for Computational Linguistics; 2005.