Named Entity Recognition on Indonesian Microblog Messages

Natanael Taufik, Alfan F. Wicaksono, Mirna Adriani
Information Retrieval Lab.
Faculty of Computer Science
Universitas Indonesia
Depok, Republic of Indonesia
{natanael.taufik}@ui.ac.id, {alfan, mirna}@cs.ui.ac.id

Abstract—**This paper describes a model to address the task of named-entity recognition on Indonesian microblog messages due to its usefulness for higher-level tasks or text mining applications on Indonesian microblogs. We view our task as a sequence labeling problem using machine learning approach. We also propose various word-level and orthographic features, including the ones that are specific to the Indonesian language. Finally, in our experiment, we compared our model with a baseline model previously proposed for Indonesian formal documents, instead of microblog messages. Our contribution is two-fold: (1) we developed NER tool for Indonesian microblog messages, which was never addressed before, (2) we developed NER corpus containing around 600 Indonesian microblog messages available for future development.**

I. Introduction

Social media is becoming an integral part of life as the use of social media is engrained in our daily activities, such as sharing information, looking for product reviews, making decisions by seeing public opinions, etc. Furthermore, business world also thinks that social media is an indispensable tool to support marketing, promote brands, and connect to online customer [1].

Social media itself has many forms, including microblog which allows users to easily share information in a relatively short sentence, a photograph, a link to a website, or a video [2]. One of the famous microblog websites is Twitter, which has gained popularity since it was established in 2006. Furthermore, Hachman [3] reported that the size of Tweet corpus has exceeded the size of collections in the U.S. Library of Congress and is still growing now. Microblog textual messages usually contain peoples opinions, experiences, tips, or reports towards events. This fact implies that the amount of useful information we can extract from microblog is very huge.

Due to huge volume of messages, many text mining tasks leverage microblog in order to build several useful applications, such as aspect-based sentiment analysis [4], [5]. Most of the tasks require good performance of several NLP tools, including named-entity recognition (NER). Unfortunately, standard NER most likely performs worse on microblog messages than formal texts since microblog messages are mostly written in non-stardard fashion. This paper addresses the task of named-entity recognition on Indonesian microblog. Microblog websites are very famous in Indonesia. In 2012, Semiocast even had mentioned that Jakarta, the capital city of Indonesia, is the most active city in terms of Twitter usage [1].

In our work, we only focus on extracting People, Location, and Organization entity types. One of the reasons is that People and Location entity types are frequent on microblog messages [6]. For Organization entity type (e.g. companies, banks, manufacturers, brands, political movements, public organizations, etc.), we really need this type for our future development related to aspect-based sentiment analysis. Furthermore, we formulate the problem as a sequence labeling problem using conditional random field model [7]. For features, we propose various linguistic, orthographic, and lookup list features, such as part-of-speech informations, last 3 letters, positions of words in a message, etc. Finally, we also developed Indonesian NER corpus containing around 600 instances, which will be available for future development and research.

II. Related Work

Named-entity recognition has been a well-studied NLP task since it was first introduced at MUC-6 [8]. Liu et al. [9] mentioned that there are three approaches for named-entity recognition: rule-based approach, machine learning approach, and hybrid approach. However, the most popular approach is certainly machine learning approach using sequence labeling algorithms, such as Hidden Markov Models (HMMs) [10] and Conditional Random Fields (CRFs) [11]. After that, further improvements were made, like the one did by Finkel et al. [12] which incorporate non-local dependencies to the sequence model. For Indonesian language, Wahyudi [13] developed NER for Indonesian formal text leveraging morphological and part-of-speech information.

The aforementioned works were mainly addressed for formal textual documents. However, microblog messages need special treatments since they are usually written in an informal way. Ritter et al. [6] reported that ”off-the-shelf” news-trained NLP tools, like POS Tagger, underperformed on Twitter messages, which can harm the performance of NER on Twitter messages. Several works have been conducted to address NER on microblog messages, such as the one that uses distant supervision [6] and the one that uses random walk model for unsupervised NER [14]. Unfortunately, until this work has been done, we did not find any work that addresses NER on Indonesian microblog messages.

[1] http://semiocast.com/en/publications/ 2012_07_30_Twitter_reaches_half _a_billion_ accounts_140m_in_the_US

978-1-5090-0922-0/16/$31.00 ©2016 IEEE

TABLE I. THE LABELS FOR SEQUENCE LABELING PROBLEM

Label	Remarks
B-PER	The beginning of a person name
I-PER	Part of a person name (except at the beginning of a person name)
B-LOC	The beginning of a location name
I-LOC	Part of a location name (except at the beginning of a location name)
B-ORG	The beginning of an organization name
I-ORG	Part of an organization name (except at the beginning of an organization name)
O	Not named-entity

TABLE II. AN EXAMPLE OF ANNOTATED INDONESIAN MESSAGES. IN ENGLISH, IT MEANS ”SUSILO BAMBANG YUDHOYONO, THE CHAIRMAN OF DEMOCRATIC PARTY, VISITED WEST JAVA”

Message	Label
Susilo	B-PER
Bambang	I-PER
Yudhoyono	I-PER
,	O
ketua	O
Partai	B-ORG
Demokrat	I-ORG
,	O
mengunjungi	O
Jawa	B-LOC
Barat	I-LOC

III. METHODOLOGY

To develop a named-entity recognizer on Indonesian microblog messages, we employed a well-known Conditional Random Field (CRF) [7], which has been shown to be very effective for the same task on English documents [11]. It means that we view our problem as a sequence labeling problem. That is, given a message containing N words $w = (w_1, w_2, ..., w_N)$, we want to find the best sequence of labels $y = (y_1, y_2, ..., y_N)$, in which each label is determined using probabilities

$$P(y_i | w_{i-l}, ..., w_{i+l}, y_{i-l}, ..., y_{i+l})$$

where l is a small number. The effectiveness of CRFs is due to the fact that they can naturally leverage information shared between neighboring positions, which is very critical for a named-entity recognition. For example, in the message ”sheffield wednesday won against manchester united !”, if we independently handle each word in the message, we will mistakenly view ”wednesday” as a day name (just like monday, tuesday, etc.). But, when we notice that ”wednesday” is preceded by ”sheffield”, it is clear that ”sheffield wednesday” is a name of a football club. Moreover, we use seven labels to indicate the entity-type that corresponds to each word in the sentence. These labels can be seen in Table I. Finally, Table II describes an example of Indonesian microblog message that has been annotated using the aforementioned labels.

Due to the learning algorithm, we certainly need to develop training corpus and devise discriminative features that can characterize the entity type of each word in the message. In this section, we present all features that we proposed for our problem. Formally, there are nine feature functions, $f_1, f_2, ..., f_9$, which map a word to a particular feature value.

1) Word. The sequence of characters makeup a word is a sufficient information to determine the entity type of a word. Suppose, f_1 denotes this feature function. Thus, the feature value for the word ketua is $f_1(ketua) = ”ketua”$.

2) Last 3 letters. This feature was inspired by the fact that Indonesian terms tend to have similar ”class” when they have the same last three letters. For example, we can easily find many Indonesian people names, like ”Setiawan”, ”Hendrawan”, ”Himawan”, etc. These terms share the same last three letters, i.e. ”wan”. For example, $f_2(Susilo) = ”ilo”$.

3) Word length. This feature has a numeric value corresponding to the number of characters in a word. For example, $f_3(Susilo) = ”wordLength : 6”$.

4) Pattern function. This feature is an adaptation from one of the features introduced by Collins [15]. It will translates a word into a special pattern using several rules: (1) All uppercased letters are mapped into letter ’A’, (2) All lowercased letters are mapped into letter ’a’, (3) All numbers are mapped into letter ’0’, (4) All other symbols are mapped into letter ’-’. There are two types of pattern can be produced, normal and summarized pattern. The difference between these two patterns is summarized pattern will reduce multiple consecutive occurence of same translated letter into one. For this feature, we use both normal and summarized pattern. For example, $f_4(Demokrat) = \{$Aaaaaaaa, Aa$\}$.

5) Inside bracket. This feature is a boolean feature with the value true if the word is located inside a bracket and false otherwise. In the sentence described in Table II, $f_5(ketua) = FALSE$.

6) Part-of-speech. Part-of-speech of a particular word and its neighboring words. Based on our experiments, the best configuration is when we leverage part-of-speech information of two previous words. To get part-of-speech of each word, we use Stanford Log-linear Part-Of-Speech Tagger[2] which are trained using manually tagged Twitter messages made by Canggadibrata and Bressan [16]. In the sentence described in Table II, $f_6(Yudhoyono) = \{$1stLeftPOS-NNP, 2ndLeftPOS-NNP$\}$.

7) Surrounding words. This feature use neighboring words as features. From our experiments, the best configuration is when we use two previous words. From Table II, $f_7(Yudhoyono) = \{$1stLeftWord-Bambang, 2ndLeftWord-Susilo$\}$.

8) Lookup list. This feature yields true if the word exactly matches at least one element of a pre-existing list and false otherwise. We currently use a manually constructed dictionary containing list of common locations as well as stopword list. Suppose, we can find the word ”Jawa” in our lookup list, then $f_8(Jawa) = TRUE$.

9) Non-standard word list. To deal with non-standard terms (e.g. slank words) appearing in microblog messages, we use a feature function that maps a word to its normal/standard form if it is found in the pre-existing list developed by Vania et al. [17]. As an example, the word ”nggak” usually appears frequently in Indonesian microblog messages. This word is actually an informal form of the word ”tidak”, which means ”no” or ”not” in English. In this case, $f_9(nggak) = ”tidak”$.

[2] http://nlp.stanford.edu/software/tagger.shtml

TABLE III. LABEL DISTRIBUTION OF OUR MANUALLY ANNOTATED TWEET MESSAGES

Label	Count
PER	225
LOC	257
ORG	204

TABLE IV. THE PERFORMANCE OF OUR BASELINE METHODS

Named-entity	Precision(%)	Recall(%)	F1(%)
PER	42.45	20.00	27.19
LOC	56.33	34.63	42.89
ORG	29.38	25.49	27.30

From example in Table II, the word Yudhoyono will has the following feature values: $f(Yudhoyono) = \{$Yudhoyono, ono, wordLength:9, Aaaaaaaaa, Aa, 1stLeftPOS-NNP, 2ndLeftPOS-NNP, 1stLeftWord-Bambang, 2ndLeftWord-Susilo$\}$.

IV. EVALUATIONS AND RESULTS

A. Data Collection

To develop training corpus, we collected Twitter messages (a.k.a Tweets) during 12 days (10th - 21st February 2015) using Twitter streaming API. Furthermore, in order to obtain many messages that are of interest, we used several keywords that are mostly Indonesian prepositions, city names, public figures, etc. Finally, we randomly selected 600 distinct Twitter messages from the collection and manually annotated those messages. After annotation, we have 379 messages which has one or more name entity tag in it. The distribution of name entity tag can be seen in Table III.

B. Results

For our experiments, we employed 10-fold cross validation on our dataset. Furthermore, precision, recall, and F1-measure were used for our evaluation metrics.

Baseline. For our baseline, we implemented methods proposed by Wahyudi [13] which use rule-based approach. Moreover, contextual and morphological information as well as part-of-speech were leveraged to characterize the entity type for each word. Table IV shows the results when we employed baseline methods on our dataset.

As we can see in Table IV, the baseline approach underperformed on our microblog messages since the proposed rules were specifically devised for formal text, instead of microblog messages. The performance on detecting person and organization names is very low since it can achieve around 27% in terms of F1 score.

TABLE V. THE PERFORMANCE OF OUR PROPOSED MODEL

Named-entity	Precision(%)	Recall(%)	F1(%)
PER	79.02	29.16	40.75
LOC	88.04	66.16	75.13
ORG	80.27	44.63	56.35

TABLE VI. AVERAGE CHANGE IN PERFORMANCE WHEN USING CAPITALIZATION FEATURE

Named-entity	Precision(%)	Recall(%)	F1(%)
PER	-2.51	+0.67	+0.63
LOC	-1.01	+1.18	+0.61
ORG	-3.01	+0.38	-0.01

Table V shows the best results of our model when we use all of the proposed features. As we can see, the performance of our model is better than the one proposed by Wahyudi [13]. The F1 scores for PER, LOC, and ORG improve around 13%, 33%, and 29%, respectively. It turns out that extracting location names become easier since we harness information from region lookup list. Moreover, there was also improvement in the performance when we lowercased all words on the message. This fact is due to inconsistency of microblog users when they post messages. For example, the term Jakarta can be found in many forms, such as ”JAKARTA”, ”jakarta”, ”JaKarTa” or ”Jakarta”. As a result, when we lowercase all those forms, they share the same meaning which comes from a single term.

C. Other Analysis

During our feature selection process, we try to use capitalization as one of the features. Knowing the nature of Tweets, capitalized letters can not be used as is like the one in formal documents [6]. In order to differentiate between Tweets which capitalization can and can not be trusted, we adapt some of the rules made by Ritter et al. [6] and modify it into two major rules: (1) First word has to be capitalized, (2) Word after punctuation which normally ends a sentence(e.g. dot, question mark, exclamation mark, semicolon) has to be capitalized. We take capitalization into account only when the message satisfy both rules.

Interestingly, capitalization feature is not performing quite well, and in fact reducing overall precision(Table VI). Capitalization feature does slightly increase recall, but after looking at overall performance, we conclude that current rules to determine capitalization is not usable yet and need improvement.

V. CONCLUSION

We have proposed a model to address the task of named-entity recognition on Indonesian microblog tweets. Recognizing named-entities on microblog messages is basically challenging due to the fact that microblog messages are usually very short and written in a non-standard way, as opposed to formal texts. Our model is based on sequence labeling task that employs Conditional Random Fields as our machine learning algorithm. Several features were proposed, including Indonesian language specific features, such as last three words and lookup list features for handling slank words. Although our model outperforms the baseline model, further improvements are needed since the best results do not seem sufficient for higher level applications.

REFERENCES

[1] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, and T. Seymour, “The history of social media and its impact on business,” *Applied Management and Entrepreneurship*, vol. 16, 2011.

[2] A. M. Kaplan and M. Haenlein, “The early bird catches the news: Nine things you should know about micro-blogging,” *Business Horizons*, vol. 54, pp. 105–113, 2011.

[3] M. Hachman, “Humanity’s tweets: Just 20 terabytes,” *http://www.pcmag.com/article2/*, vol. Accessed 4th June 2015.

[4] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, ser. Data-Centric Systems and Applications. Springer, 2007. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-37882-2

[5] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1561/1500000011

[6] A. Ritter, S. Clark, Mausam, and O. Etzioni, “Named entity recognition in tweets: An experimental study,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1524–1534. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145595

[7] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[8] R. Grishman and B. Sundheim, “Message understanding conference-6: A brief history,” in *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, ser. COLING ’96. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996, pp. 466–471. [Online]. Available: http://dx.doi.org/10.3115/992628.992709

[9] X. Liu, S. Zhang, F. Wei, and M. Zhou, “Recognizing named entities in tweets,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 359–367. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002519

[10] D. M. Bikel, R. Schwartz, and R. M. Weischedel, “An algorithm that learns what’s in a name,” *Mach. Learn.*, vol. 34, no. 1-3, pp. 211–231, Feb. 1999. [Online]. Available: http://dx.doi.org/10.1023/A:1007558221122

[11] A. McCallum and W. Li, “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons,” in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL ’03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 188–191. [Online]. Available: http://dx.doi.org/10.3115/1119176.1119206

[12] J. R. Finkel, T. Grenager, and C. Manning, “Incorporating non-local information into information extraction systems by gibbs sampling,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL ’05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363–370. [Online]. Available: http://dx.doi.org/10.3115/1219840.1219885

[13] G. Wahyudi, “Pengenalan entitas bernama berdasarkan informasi kontekstual, morfologi, dan kelas kata (in indonesian),” *Depok, Indonesia. Universitas Indonesia*, 2004.

[14] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, “Twiner: Named entity recognition in targeted twitter stream,” in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’12. New York, NY, USA: ACM, 2012, pp. 721–730. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348380

[15] M. Collins, “Ranking algorithms for named-entity extraction: Boosting and the voted perceptron,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ser. ACL ’02. Philadelphia: Association for Computational Linguistics, 2002, pp. 489–496. [Online]. Available: http://dl.acm.org/citation.cfm?id=1073165

[16] H. F. Canggadibrata and S. Bressan, “Part of speech tagging for microblogging posts in indonesian,” in *The 6th International Workshop on Malay and Indonesian Language Engineering 2012*, ser. MALINDO ’12. MALINDO, 2012, pp. 14–28.

[17] C. Vania, M. Ibrahim, and M. Adriani, “Sentiment lexicon generation for an under-resourced language,” *International Journal of Computational Linguistics and Applications (IJCLA) (To Appear)*, 2014.