Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 221 – 228

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning

Aditya Satrya Wibawa*, Ayu Purwarianti

School of Electrical Engineering and Informatics, Institut Teknologi Bandung

Abstract

Here, we describe our effort in building Indonesian Named Entity Recognition (NER) for newspaper article with 15 classes which is larger number of class type compared to existing Indonesian NER. We employed supervised machine learning in the NER and conducted experiments to find the best attribute combination and the best algorithm with highest accuracy. We compared the attribute of word level, sentence level and document level. In the algorithm, we compared several single machine learning algorithms and also an ensembled one. Using 457 news articles, the best accuracy was achieved by using ensemble technique where the result of several machine learning algorithms were used as the feature for one machine learning algorithm.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Indonesian; Named Entity Recognition (NER); machine learning; ensemble

1. Introduction

The term "Named Entity" has been used widely in the field of information extraction (IE), question answering (QA), and various fields of applications of natural language processing (NLP) other[1]. The term was first used in the Message Understanding Conference-6 (MUC-6) in 1995. NERC considered as an important component of the technology that underlies many other NLP applications, such as information extraction (IE), question answering (QA), text summarization, information retrieval (IR), etc[2].

* Corresponding authors.
E-mail addresses: aditya.satrya@gmail.com

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.053

NERC research has been widely done for English because it is the most widely used language in the world. In addition, the availability of datasets in English was so high to facilitate NERC research, which generally requires a large dataset. NERC for languages other than English is also widely applied in smaller proportions. Sekine et al. NERC stated that there were researches seeking to overcome the problems of multi-language and language-independence. Moreover, German, Spanish, Dutch, Japanese, Chinese, French, Greek, Italian, Bulgarian, Catalan, Korean, Hindi, Romanian, Russian, Swedish, Turkish, Portuguese, and Arabic is the language that gets great attention and actively investigated[2].

NERC researches had been done on various types of texts and domains. Types of widely used documents are news articles, email, scientific literature, and document / religious texts. One example of a popular specific domain is the field of bioinformatics, where NERC is required to identify the type named-entity like "protein", "DNA", "RNA" of various scientific literature of biology.

NERC study on open domain for Bahasa Indonesia is still rarely done, such as is done Budi et al.[3,4] and Yanuar[5]. Therefore, this study is intended to deepen the study and experimentation with statistical methods of NERC in Indonesian on the open domain. The definition of open-domain is the widest possible coverage of domains that contain the words that can be understood easily by adults in general. Scope of the study in this research is in terms of a combination of features, learning schemes, and the selection of machine learning algorithms.

With the requirements as above, the most appropriate document is a news article because in it there is a multi-domain text by category in general news (politics, economics, entertainment, etc.). Moreover, the terms used in news articles are words that can be understood by most adults without requiring the help of a dictionary or specialized knowledge in a particular field. For the same reasons, Sekine et al. use a corpus of news articles to hierarchically categorize named-entities in open domain[6]. In addition, the news document has some profitable characteristics in NERC, such as the formal language used in the article and elements contained in the article are the formal named-entity name (who are the person/organization, what events are happening), location (where the), and temporal (when it happens)[7,8].

2. Related Works

Using statistical method for NERC includes two important things: the classification scheme and the features used. Various classification schemes have been applied in previous studies. Among them are one-phased learning scheme using various machine learning algorithms such as HMM[9], Decision Tree[10], Maximum Entropy[5,11], SVM[12,13,14], CRF[15], and Association rules mining[3]. In addition, there is a voting scheme of SVM classifier, CRF, and Maximum Entropy[16]. Hierarchical classification scheme has been undertaken against Person class in limited cases[17].

As for the employed features, it can be grouped into word-level features, sentence level features and list lookup features. The word-level features consist of lexical, POS, and morphological features. Each of features has several candidates such as the word list type, the window width or chunk label.

For Bahasa Indonesia, NERC research is relatively few. Among the NERC research in Bahasa Indonesia, two of them are researches conducted by Indra Budi, et al. In the first study, the features are combined from contextual features, morphological, and part-of-speech; and the classification was done by knowledge engineering approach, i.e. rules made by experts[4]. In the second study, association rules mining is used to identify and resolve the co-reference problems. The study employed several morphological and lexical features such as pronoun class and class name, string similarity, and the position of the text[3].

3. Named-Entity Classes

The named-entity type employed in this research is adapted from Sekine et al with slight adjustments. The list of named-entity type is described in Table 1. Adjustments made by combining the class Money, Stock-Index, Point, Percent, Multiplication, Frequency, Rank, Age, Ordinal-Number, Latitude-Longitude, and School-Age Numex into one class because of the small frequency of each class and those classes have similar patterns of occurrence each other, which appears in the combination of numbers and symbols. With the incorporation of this class, expected positive samples of these classes can be statistically significant which will impact on the increased accuracy for the combined class Numex.

Measurement and Countx class that also included into the Numex category are not joined together because both classes have typical appearance pattern and have many subclasses. For example, Measurement class has 10 subclass and has appearance pattern in the form of a number followed by a unit name like "5 km", "400 J", while Countx class has 7 subclass and the has appearance pattern of a noun followed by numbers like "5", "100 building".
Reasons for selecting hierarchy extended named-entity (ENE) Sekine is because this hierarchy is extensible for further research (e.g. classification within the hierarchy of classes that are fine-grained), minimizing ambiguity during the annotation (the GOE class and GPE), and have the domain coverage the most extensive (there are 150 NE types).

Table 1. Named-Entity Class Used in This Study

Class	Example
Person	Aditya Satrya, Susilo Bambang Yudhoyono, Mahfudz
God	Allah, Yesus Kristus
Organization	DPR, Partai Keadilan Sejahtera, Kemenkum dan HAM
Location	Indonesia, Desa Babakan, Gunung Gede
Facility	Taman Safari, Institut Teknologi Bandung, Bandara Soekarno-Hatta
Product	Toyota Camry, RAPBN 2011, UU No. 23
Event	Idul Fitri, Perang Dunia II, Tsunami Aceh
Natural-Object	Hidrogen, Elang Bondol
Disease	AIDS, Demam berdarah, Hepatitis
Color	Merah, Magenta, Biru
Timex	Pukul 15.00, 1 April 2013
Periodx	5 semester, 10 hari, 5 tahun
Numex	1.5, 5%, Rp50.000.000
Countx	10 orang, 5 rumah
Measurement	60 km, 500 J

4. Corpus

Corpus used in this research consists of 457 news articles that belong to 3 categories, which is derived from the Indonesian-language online news site in a span of 4 days. Web site used is Detik (http://detiknews.com), Kompas (http://kompas.com), and Media Indonesia (http://mediaindonesia.com).

The number of named-entity contained in the corpus is 16,738 terms such as shown in Table 2 below. 50% of the corpus is used as the training data and 50% as the testing data.

Table 2. Number of Named-Entity Terms in the Corpus

NE Class	Training Data		Testing Data	
	Count	%	Count	%
Person	1359	16%	1406	17%
God	26	0%	14	0%
Organization	1459	17%	1561	19%
Location	1819	22%	1652	20%
Facility	330	4%	338	4%
Product	1281	15%	1312	16%
Event	169	2%	156	2%
Natural-Object	69	1%	17	0%

NE Class	Training Data		Testing Data	
	Count	%	Count	%
Disease	0	0%	2	0%
Color	18	0%	4	0%
Timex	819	10%	877	10%
Periodx	107	1%	89	1%
Numex	397	5%	437	5%
Countx	429	5%	370	4%
Measurement	85	1%	136	2%
TOTAL	8367	100%	8371	100%

Corpus consists of 1,500 sentences and 126,820 tokens where 25,848 tokens having named entity class such as listed in Table 4 1 and 100,978 tokens with NONE class. The comparison of token size with NE class and Not NE class is shown in Table 3 below. The comparison shows the unbalanced data contained in the corpus.

Table 3. Comparison of the Token Size between NE class and not NE class

Token	Count	%
With NE class	25,848	20%
With NONE class	100,972	80%
Total	126,820	100%

5. Classification Feature

5.1. Word-level feature

Word-level features are derived features of a word is independent of the context. Word-level features consist of:

1. *Kind*, the type of token based on constituent character. The value consists of word, symbol, number, punctuation, URL, email, date, and time.
2. *Orthography*, the characteristic combination of uppercase and lowercase letters of the characters making up a token. The value consists of UpperInitial (starting capitalized), AllCaps (all uppercase), and MixedCaps (combination of uppercase and lowercase letters).
3. `HasPeriod`, the presence / absence of punctuation point "." in token. This feature is worth the End (at the end of the token), Internal (in the middle of the token), None (no).
4. `HasAposthrope`, the presence / absence of punctuation apostrophe in tokens (Yes/No).
5. `HasHypen`, the presence / absence of punctuation hyphen in token (Yes/No).
6. `HasDigit`, the presence or absence of numeric characters in the token (Yes/No).
7. `SummarizedPattern`, obtained by substitution type uppercase alphabetic characters replaced with "A", with a lowercase "a", and punctuation symbols "-", and the numeric character "0" then eliminate repetition of consecutive characters. For example, the character of the token "Hidayat" is substituted to "Aaaaaa", then the repetition eliminated to be "Aa".
8. *Part-of-speech* (POS). Number of POS classes used were 35[23, 24], which accuracy was 99.4% (non-OOV) and 91.3% (30% OOV).

5.2. Sentence-level features

Sentence-level feature employed in this study is the position of the token in the sentence. This feature is named positionInSentence. This feature has a nominal value of tokens that begin to be in the beginning of the sentence, and the middle to later token.

Based on observation of the data, this feature is expected to differentiate token due `UpperInitial` are at the beginning of sentences and because it must be written with a capital letter (e.g., names of people or organizations).

5.3. List-lookup features

Lookup features list is a feature derived from the results of the lookup list that has been compiled before. Compiled the following list with corresponding features among others:

1. Indonesian Dictionary, the dictionary entry in the list that begins with a lowercase letter.
2. *Stop Word*, the list of entries in the dictionary that includes stopword. List taken from the lemma in the dictionary that has a class of words and pronouns.
3. *Capitalized Noun*, the lemma list is written with capital letters in the dictionary.
4. *Named-entity list*, the list of named-entity for each class. Lists obtained from Wikipedia website.
5. *Clue list*, the list of words that can be used as guidelines in detecting and classifying named-entity. List of manually constructed through observation of the corpus, and the named-entity list

5.4. Contextual features

We employed 3 preceding words and 3 succeeding words as the word window of contextual features. The complete features for the word window are the lexical feature, the sentence level lookup lists, and the NE class (for the preceding words).

6. Experiments

6.1. NERC Architecture

NERC processes the un-annotated document into an NE-annotated document through a series of processes, i.e. tokenization, sentence splitting, word-level feature extraction, POS tagging, sentence-level feature extraction, lookup lists, and classification. The processing workflow is illustrated in Fig 1.

6.2. Experiment Scenario

The scenario was prepared to consider some aspects. First aspect is the stages of classification. This scenario aims to determine which scheme best classification stages. Based on the stage, the classification is divided into:

1. **Direct**. Classification is done to predict the class of named-entity tokens into 15 classes of named-entity level 1 plus 1 NONE class. Thus, there were 16 target classes, namely Person, God, Organization, Location, Facility, Product, Event, Natural-Object, Disease, Color, Timex, Periodx, Numex, Countx, Measurement, and NONE.
2. **Incremental**. This classification is done through 2 stages:
   a. ***Stage I.*** Classification is done to predict whether a token is part of a named-entity or not. So, there are only 2 pieces of the target class, i.e. NE (part of named-entity) and NONE (outside of named-entity).
   b. ***Stage II.*** Classification is done to predict the token in Phase I as a named entity. There are 15 pieces of the target classes, the same class with Direct classification reduced class NONE because the class is classified in stage I.

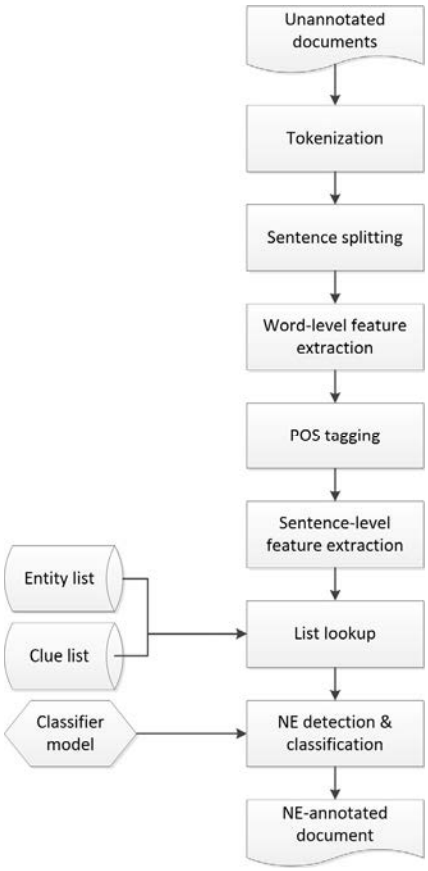

Fig 1. NERC Architecture

Second, it considers the combination of features. Classification is done by combining various features of word level, sentence level, lookup lists, and contextual features. This scenario aims to determine the significance of the various features. Based on its combination, the classification is divided into:

1. **W+S.** Classification is done by combining the features of word-level (lexical), sentence level, and contextual features. Word level features and sentence level features are combined because the sentence level consists of only one feature, namely `positionInSentence`.
2. **W+S+L.** Classification is done by combining the features of word-level (lexical), sentence level, lookup lists, and contextual features.

By considering aspects of the classification stages and a combination of features, structured testing scenario consists of 6 test cases. Matrix of test cases is included in Table 4.
Third, aspects of the algorithms used. The algorithms to be compared are Naïve Bayes, SVM[18], and Simple Logistic[19].

Table 4. Test Scenario

Classification stage/ Feature combination	W+S	W+S+L
Direct	Direct; W+S	Direct; W+S+L
Stage-I	Stage-I; W+S	Stage-I; W+S+L
Stage-II	Stage-II; W+S	Stage-II; W+S+L

6.3. Tools

Tool used as a text processing framework is GATE 6.0 build 3764[21], while tool used for machine learning algorithms is WEKA 3.6.4[22]. GATE stands for General Architecture for Text Engineering. GATE can be used for almost all text processing by providing a document format, user interface, and a wrapper for various text processing algorithms.

The WEKA stands for Knowledge Analysis Waikato Environment, is an integrated workbench that contains a collection of machine learning techniques. This study integrates WEKA classifier models into the GATE architecture.

7. Conclusion

In this research, a statistical based NERC has been conducted to Indonesian news documents. Several things can be concluded from the experiments.

Features used can be grouped into 3 categories, those are word-level (morphological and POS), sentence-level, and the lookup list. Use of word-level features can be considered enough for classification into ENE classes and NONE, although if lookup feature is added, it will improve the accuracy. The best performance results on testing is at 0.528 of F-measure generated by the algorithm Simple Logistic, Direct scheme, and features combination of word-level, sentence-level, and the lookup list.

Table 5. Test results of Direct classification

Algorithm	Feature Level	Baseline			Chunking			Baseline + Feature Selection			Chunking + Feature Selection		
		Prec	Rec	F	Prec	Rec	F	Pre	Rec	F	Pre	Rec	F
Naïve Bayes	W+S	0.349	0.414	0.379	0.277	0.412	0.332	0.434	0.426	0.430	0.254	0.302	0.276
	W+S+L	0.386	0.582	0.464	0.363	0.577	0.445	0.501	0.583	0.539	0.488	0.589	0.534
SVM	W+S+L	0.389	0.483	0.431	0.411	0.444	0.427	0.482	0.529	0.504	0.495	0.488	0.491
Simple Logistic	W+S+L	0.481	0.515	0.497	0.441	0.447	0.444	0.513	0.544	0.528	0.495	0.489	0.492

Table 6. Test results of Incremental classification

Algorithm	Feature Level	Baseline			Chunking			Baseline + Feature Selection			Chunking + Feature Selection		
		Prec	Rec	F	Prec	Rec	F	Pre	Rec	F	Pre	Rec	F
Naïve Bayes	W+S	0.405	0.429	0.416	0.382	0.412	0.397	0.388	0.418	0.402	0.314	0.339	0.326
	W+S+L	0.521	0.285	0.369	0.522	0.282	0.366	0.52	0.285	0.368	0.52	0.283	0.367
SVM	W+S+L	0.453	0.49	0.471	0.437	0.473	0.454	0.485	0.525	0.504	0.462	0.501	0.481
Simple Logistic	W+S+L	0.478	0.518	0.497	0.453	0.491	0.471	0.475	0.515	0.494	0.473	0.513	0.492

8. Acknowledgment

This work is partially supported by the Indonesia Endowment Fund of Education (LPDP), Ministry of Finance, Indonesia.

References

1. S. Sekine, "Named Entity : History and Future," 2003.
2. D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investigationes, vol. 30, no. 1, p. 3–26, 2007.
3. I. Budi and S. Bressan, "Application of association rules mining to Named Entity Recognition and co-reference resolution for the Indonesian language," International Journal of Business Intelligence and Data Mining, vol. 2, no. 4, p. 426–446, 2007.

4. I. Budi, S. Bressan, G. Wahyudi, Z. Hasibuan and B. Nazief, "Named entity recognition for the Indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach," Discovery Science, pp. 57-69, 2005.
5. M. R. Yanuar, Adaptasi Named-Entity Recognizer pada OpenNLP untuk Pemroses Bahasa Indonesia, Bandung: Institut Teknologi Bandung, 2013.
6. S. Sekine, K. Sudo and C. Nobata, "Extended named entity hierarchy," in Proceedings of LREC, 2002.
7. R. M. S. Putra, Teknik Menulis Berita dan Feature, 2006 ed., Jakarta: PT. INDEKS Kelompok Gramedia, 2006.
8. M. Mohd, "Named Entity Patterns across News Domains," BCS IRSG Symposium: Future Directions in Information Access, pp. 30-36, 2007.
9. D. M. Bikel, S. Miller, R. Schwartz and R. Weischedel, "Nymble : a High-Performance Learning Name-finder," 1991.
10. S. Sekine, "NYU: Description of the Japanese NE system used for MET-2," in Proc. of the Seventh Message Understanding Conference, 1998.
11. A. Borthwick, J. Sterling, E. Agichtein and R. Grishman, "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition," in Proc. of the Sixth Workshop on Very Large Corpora, 1998.
12. M. Asahara and Y. Matsumoto, "Japanese Named Entity Extraction with Redundant Morphological Analysis Masayuki," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003.
13. A. Ekbal and S. Bandyopadhyay, "Named entity recognition using support vector machine: A language independent approach," Int. J. Elec. Comput. Syst. Eng, vol. 4, p. 155–170, 2010.
14. O. S. Kuspriyanto, D. H. Widyantoro and H. S. Sastramihardja, "Performance Evaluation of SVM-Based Information Extraction using τ Margin Values," International Journal on Electrical Engineering and Informatics, vol. 2, no. 4, p. 256–265, 2010.
15. A. Ekbal, R. Haque and S. Bandyopadhyay, "Named entity recognition in Bengali: A conditional random field approach," in Proceedings of IJCNLP, 2008.
16. A. Ekbal and S. Bandyopadhyay, ""“Improving the Performance of a NER System by Post-processing and Voting," Structural, Syntactic, and Statistical Pattern Recognition, p. 831–841, 2010.
17. A. Ekbal, E. Sourjikova, A. Frank and S. P. Ponzetto, "Assessing the Challenge of Fine-Grained Named Entity Recognition and Classification," ACL 2010, no. July, p. 93, 2010.
18. J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," 1999.
19. M. Summer, E. Frank and M. Hall, "Speeding up Logistic Model Tree Induction," Knowledge Discovery in Databases: PKDD 2005, pp. 1-8, 2005.
20. R. B. Grishman and Sundheim, "Message understanding conference-6: A brief history," in Proceedings of the 16th conference on, 1996.
21. H. Cunningham, Y. Wilks and R. J. Gaizauskas, "GATE - a General Architecture for Text Engineering," Intelligence, p. 1057–1060, 1995.
22. M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software : An Update," SIGKDD Explorations, vol. 11, no. 1, pp. 10-18.
23. A. F. Wicaksono and A. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," in Proceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop, 2010.
24. A. Purwarianti, A. Saelan, I. Afif, F. Ferdian, A. F. Wicaksno, "Natural Language Understanding Tools with Low Language Resource in Building Automatic Indonesian Mind Map Generator, ", International Journal on Electrical Engineering and Informatics, vol. 5, no. 3, pp. 256-269, 2013.