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.