NAMED ENTITY TAGGING FOR INDONESIAN DOCUMENTS USING RULE-BASED METHOD

Pangudi Citraningputra, Julio Adisantoso

Department of Computer Science

Faculty of Mathematics and Natural Sciences

Bogor Agricultural University

julioipb@gmail.com

2010

Abstract

The main objective of this research is to implement named entity tagging in Indonesian documents. Rule-based method is used for this research. This method uses various rules to identify words or phrase to know if that words or phrase can be classified as named entity. Named entity form obtained from the named entity tagging can be classified in six forms, there are {NAME}, {ORGANIZATION}, {LOCATION}, {CURRENCY}, {DATE}, {TIME}, and {NUMBER} to identify the name of people, organizations, places, currency, date, time, and number. In addition of using rules, dictionary is also used to identify named entity {NAME}, {ORGANIZATION}, and {LOCATION}.

The evaluation is done by using 91 document samples from a total of 1.000 documents. Evaluation is based on rules and performance of the system. Evaluation based on the rule is done by comparing the result from the manual test with the result from the system test. The results of the evaluation can identify the accuracy level of the system. There are 3.599 named entities classificated from the total 3.641 named entities in 91 documents. There are 99 unclassified named entities, 87 wrong classified named entities. Entities are classified consist of 514 NAME named entities, 576 ORGANIZATION named entities, 1.376 LOCATION named entities, 117 CURRENCY named entities, 341 DATE named entities, 4 TIME named entities, and 680 NUMBER named entities. From this evaluation, it can be concluded that the more documents are analyzed, the better level of identification can be found. In the evaluation of system performance, evaluation is done by the time and the number of words, and by the time and the many of named entities. From this evaluation, it also concluded that the number of words and the many of named entities contained in the document is very influential on the time required in the process of named entity tagging.