Pembangunan Dataset NE Bahasa Indonesia dari Data Wikipedia dan DBpedia dengan Metode *Entities Expansion* pada DBpedia

Haji Dito Murya Alfarohmi[1], Moch. Arif Bijaksana[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]ditomry@student.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id

Abstract

In Indonesian, the NER (Named Entity Recognition) system still needs a lot of improvement. Though NER is a major component in IE (Information Extraction) which is used by other advanced components. To create a reliable Indonesian NER system using a machine learning approach, large dataset are needed. If the dataset is constructed by tagging it manually, the size of generated dataset is very small. Therefore, a system was created to build Indonesian NE (Named Entities) dataset which were tagged automatically using Wikipedia as a source of corpus and DBpedia as NE labeling references with the Entities Expansion method to expand DBpedia labeling references NE. Currently the existing system cannot detect names containing words beginning with lowercase letters on automatic tagging, have not tried adding gazetteers to person entity, and the DBpedia Entities Expansion method rules can still be modified to produce better NE labeling reference quality. In this final project a system was built to overcome these shortcomings. Evaluation shows, the best Indonesian NE dataset built in this final project produces F1-score of 54.93%, 3.32% higher than the results of previous studies 51.61%. This best dataset is built by adding a detection method to automatic tagging, using DBpedia Entities Expansion modifications, but without adding gazetteers to person entity.

Keywords*: *Wikipedia, DBpedia, Entities Expansion, Automatically Tagging, Indonesian NE Dataset