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

Abstrak

Pada bahasa Indonesia, sistem NER (*Named Entity Recognition*) masih memerlukan banyak perbaikan. Padahal NER adalah komponen utama dalam IE (*Information Extraction*) yang digunakan oleh komponen lanjutan lainnya. Untuk menciptakan sistem NER bahasa Indonesia yang andal menggunakan pendekatan *machine learning*, diperlukan dataset yang besar. Apabila dataset dibangun dengan melabeli secara manual, ukuran dataset yang dihasilkan sangat kecil. Oleh sebab itu, dibuat sistem untuk membangun dataset NE (*Named Entities*) bahasa Indonesia yang dilabeli secara otomatis menggunakan data Wikipedia sebagai sumber korpus dan DBpedia sebagai referensi pelabelan NE dengan metode *Entities Expansion* untuk memperluas referensi pelabelan NE DBpedia. Saat ini sistem yang ada belum dapat mendeteksi nama yang mengandung kata diawali huruf kecil pada pelabelan otomatisnya, belum mencoba menambahkan *gazetteers* entitas *person*, serta aturan metode DBpedia *Entities Expansion* masih dapat dimodifikasi untuk menghasilkan kualitas referensi pelabelan NE yang lebih baik. Pada tugas akhir ini dibangun sistem yang mengatasi kekurangan tersebut. Evaluasi menunjukkan, dataset NE bahasa Indonesia terbaik yang dibangun pada tugas akhir ini menghasilkan F1-*score* 54,93%, lebih tinggi 3,32% dari hasil penelitian sebelumnya 51,61%. Dataset terbaik ini dibangun dengan menambahkan metode deteksi pada pelabelan otomatis, menggunakan DBpedia *Entities Expansion* modifikasi, tetapi tanpa menambahkan *gazetteers* entitas *person*.

Kata kunci : Wikipedia, DBpedia, *Entities Expansion*, Pelabelan Otomatis, Dataset NE Bahasa Indonesia