PROPOSAL PROYEK

11S4037 – Pemrosesan Bahasa Alami

Named Entity Recognition of Indonesian Text

Disusun oleh:

12S18020 Dita L. Sastri Sihombing
12S18029 Estomihi Rascana Sirait
12S18061 Angela Friscilia Simamora

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

2021

DAFTAR ISI

DAFTAR TABEL

BAB I PENDAHULUAN

Pada bab ini menyajikan latar belakang, tujuan, manfaat, dan ruang lingkup pengerjaan proyek.

1.1 Latar Belakang

Named Entity Recognition merupakan kata benda yang mengacu pada jenis individu tertentu seperti nama organisasi, nama orang, nama lokasi, dan sebagainya [1]. *Named Entity Recognition* melibatkan pemrosesan teks dan mengidentifikasi kemunculan kata atau ekspresi tertentu sebagai milik kategori tertentu *Named Entity* (NE). Perangkat lunak pengenalan *named entity* berfungsi sebagai alat pemrosesan awal yang penting untuk tugas-tugas seperti ekstraksi informasi, pengambilan informasi, dan aplikasi pemrosesan teks lainnya. Apa yang dianggap sebagai *named entity* bergantung pada aplikasi yang menggunakan anotasi. Salah satu aplikasi tersebut adalah pengambilan dokumen atau penerusan dokumen otomatis: dokumen yang dicatat dengan informasi *named entity* dapat dicari lebih akurat daripada teks mentah [2].

Dalam beberapa dokumen teks untuk memperoleh banyak informasi yang penting seperti nama orang, nama lokasi, nama organisasi yang dimana dalam dokumen tentu dilakukan dengan manual yaitu membaca keseluruhan teks yang ada, diperlukan waktu yang banyak lagi jika sebuah dokumen sangat panjang. Di saat sekarang ini sudah banyak dilakukan pembahasan terkait *Named Entity,* mengingat bahwa entitas dari sebuah dokumen itu penting dan dengan upaya yang dilakukan membuat *Named Entity Recognition* dapat digunakan untuk mendeteksi informasi secara otomatis sehingga tidak perlu menghabiskan banyak waktu untuk membaca dokumen teks keseluruhan. *Named Entity Recognition* dapat diimplementasikan pada machine translation, question answering dan semantic web.(Leonandya, 2015) [1]. Namun dalam beberapa sumber penelitian terdapat kekurangan pada proses *automatic tagging*. *Automatic tagging* merupakan proses untuk melakukan *tagging* pada setiap kata atau frasa dengan jenis entitasnya. Maka dari itu dalam proyek ini mencoba meningkatkan kemampuan *automatic tagging* dalam mengimplementasikan *POS-Tagging* dengan beberapa aturan tambahan pada proses automatic tagging tersebut. Dengan mengimplementasikan *POS-Tagging* dengan tujuan untuk memperoleh seluruh kata atau frasa yang memiliki kemungkinan mempunyai jenis entitas dan selanjutnya kata atau frasa tersebut akan dilakukan pengecekkan dengan *rule* yang telah disediakan dan akan dilakukan tag entitas berdasarkan aturan yang digunakan.

1.2 Rumusan Masalah

Berdasarkan pembahasan masalah yang telah dibahas sebelumnya, dimana terdapat kekurangan pada saat proses *automatic tagging* pada pelabelan banyak kata atau frasa. Maka dari itu, diperlukan metode yang diperlukan untuk meningkatkan proses kerja *automatic tagging* dalam melakukan pelabelan.

1.3 Tujuan

Tujuan proyek ini adalah:

1. Menghasilkan model *Named Entity Recognition* dengan *tag entitas* pada kata dalam dokumen teks Bahasa Indonesia.
2. Memperkaya pengetahuan bagi penulis maupun pembaca terkait *Named Entity Recognition* dengan *POS-Tagging.*

1.4 Manfaat

Manfaat proyek ini adalah:

1. Memberikan model *Named Entity Recognition* dengan tag entitas yang sesuai pada kata atau frasa yang terdapat dalam dokumen teks Bahasa Indonesia dengan memanfaatkan pemrosesan bahasa alami.
2. Sebagai tahapan awal dalam *Information Extraction* Bahasa Indonesia

1.5 Ruang Lingkup

Ruang lingkup dari proyek ini adalah:

1. Membangun sistem *Named Entity Recognition* yang digunakan untuk menganalisis teks Bahasa Indonesia
2. Menggunakan set data berisi teks Bahasa Indonesia, yaitu teks SINGGALANG (https://github.com/ialfina/ner-dataset-modified-dee/tree/master/singgalang)

1.6 Sistematika Penyajian

Adapun sistematika penyajian dari proyek ini adalah:

Bab 1. Pendahuluan, membahas tentang latar belakang, pertanyaan penelitian, tujuan, manfaat, ruang lingkup, dan sistematika penyajian penelitian.

Bab 2. Isi, menjelaskan mengenai teori-teori yang berkaitan dengan proyek diantaranya metode yang digunakan, proses dan perangkat (tools) terkait dengan tujuan penelitian.

Bab 3. Rencana, menjelaskan terkait jadwal kerja pengerjaan proyek dan pembagian tugas.

BAB II ISI

Pada bab ini dijelaskan tahapan pemrosesan bahasa alami yang akan diterapkan.

2.1 Perumusan Masalah

Pada tahap ini dilakukan identifikasi masalah yang terjadi pada objek penelitian yang juga dapat dirumuskan sebagai tujuan dari penelitian proyek.

2.2 Pengumpulan Data

Pada tahap ini dilakukan mengumpulkan data-data yang diperlukan dalam pengerjaan proyek dalam menyelesaikan masalah.

2.3 Perancangan Sistem

Pada tahap ini dijelaskan metode yang akan digunakan dalam merancang sistem dalam melakukan pengecekan bahasa pada teks dokumen.

2.4 Training

Untuk memperoleh data training yang lebih baik, Penulis menggunakan POS-Tagging untuk meningkatkan kualitas dari data training. POS-Tagging digunakan untuk mengetahui kelas kata masing-masing kata dimana kelas kata ini kemudian akan digunakan untuk menentukan kata-kata yang memiliki entitas *person*, *location* atau organization.

2.5 Testing

Untuk menguji kemampuan model hasil training, Penulis melakukan pengujian pada teks Bahasa Indonesia “SINGGALANG”. Pengujian dilakukan dengan mengambil kalimat secara acak dan dilakukan tag entitas pada kalimat-kalimat tersebut secara manual

2.6 Evaluasi

Pada tahap ini dilakukan analisis terhadap hasil akhir dari proses POS-Tagging dalam mengklasifikasi kata atau frasa pada teks Bahasa Indonesia serta untuk mengukur keakuratan sistem.

2.7 Kesimpulan dan Saran

Pada tahap ini menyimpulkan hasil proyek yang telah dilakukan berdasarkan hasil pengolahan data. Dari hasil evaluasi proyek juga diberikan saran yang berkaitan dengan proses pengerjaan agar dapat dilakukan lebih baik lagi pada proyek selanjutnya.

BAB III RENCANA KERJA

Pada bab ini akan dijelaskan jadwal kegiatan dan pembagian tugas.

3.1 Jadwal Kegiatan

Jadwal Kegiatan direncanakan akan dilaksanakan selama 4 minggu. Tahapan dimulai dari perumusan masalah hingga evaluasi proyek. Rincian kegiatan lebih jelas dapat dilihat pada Tabel 1.

Tabel 1 Jadwal Kegiatan

No	Kegiatan	Minggu																									
		Minggu 1							Minggu 2							Minggu 3							Minggu 4						
		1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	Perumusan Masalah																												
2	Pengumpulan Data																												
3	Perancangan Sistem																												
4	Training																												
5	Testing																												
6	Evaluasi																												
7	Kesimpulan dan Saran																												

3.2 Pembagian Tugas

Pembagian tugas akan dilampirkan pada laporan akhir.

DAFTAR PUSTAKA

[1] A. WILLYAWAN, "NAMED ENTITY RECOGNITION (NER) BAHASA INDONESIA," p. 54, 2018.

[2] M. M. d. C. G. Andrei Mikheev, "Named Entity Recognition without Gazetteers," p. 8, 1999.