

BAB I

PENDAHULUAN

1.1. Latar Belakang

Natural language processing merupakan sebuah interdisiplin ilmu dari disiplin *machine learning* yang dapat memproses dan mengeluarkan hasil berupa bahasa natural[1]. Dalam pemrosesan *natural language processing* diperlukan beberapa tahapan yang dilakukan sebelum mendapat hasil akhir berupa teks natural, salah dua tahapan itu merupakan *dependency parser* dan *named entity recognition*[1]. *Dependency parser* merupakan tahapan dimana mesin menganalisis arti dari sebuah bahasa manusia dengan cara membandingkan hubungan arti dari satu kata ke kata yang lain dalam sebuah kalimat[1], [2]. Teknik *dependency parser* sendiri merupakan salah satu dari 2 teknik utama dalam *text parsing*: *constituent parsing* dan *dependency parser*[2]. Dalam pemrosesan teks bahasa Indonesia *dependency parser* lebih cocok digunakan dibanding *constituent parser* karena sifatnya yang cocok untuk bahasa tanpa struktur *grammar*[2]. *Named entity recognition* merupakan teknik untuk mengekstrak informasi dalam teks ke dalam entitas tertentu seperti lokasi, nama orang, atau organisasi[3].

Natural language processing dalam perkembangannya mengalami peningkatan riset terutama dalam penggunaan di bidang industri dan perusahaan terutama industri media massa dan studi bahasa. Penggunaan teknologi *natural language processing* tentu dapat membantu pekerjaan manusia seperti halnya teknologi lain. Riset untuk teknologi ini di Indonesia sendiri masih sedikit[1]. Terutama pada *library spacy* salah satu *library* yang cukup terkenal di dalam dunia *natural language processing*. Bahasa Indonesia sendiri di dalam *library spacy* belum mencapai tahap rilis di dalam repositorinya dikarenakan model dan *file – file* penunjangnya belum lengkap. *Dependency parser* dan *named entity recognition* menjadi salah satu model yang belum ada di dalam *library spacy*. Dikarenakan penggunaan *natural language processing* sudah mulai banyak

dipakai perusahaan di Indonesia, maka pembuatan skripsi ini diharapkan dapat membantu pengguna *library spacy* di Indonesia yang ingin membuat proyek atau aplikasi *natural language processing*.

Model yang akan dibuat adalah salah dua teknik dalam memproses sebuah bahasa. Pembuatan model ini akan melibatkan *dataset* yang cukup besar dan banyak. Model ini akan dibuat dengan menggunakan *library spacy* seutuhnya yang menggunakan algoritma *convolutional neural network*. Pembuatan model ini akan menggunakan *spacy cli* untuk *training* dan *evaluate*, dan *jupyter notebook* untuk *scripting* dikarenakan tampilan dan *workflow*-nya yang dinamis dan mudah dipahami seperti membaca buku. Diharapkan dengan pembuatan model ini akan membuat teknologi *natural language processing* makin mudah dikerjakan di Indonesia dan juga dapat dipelajari lebih lanjut ketika riset *natural language processing* makin banyak dikerjakan dan dipergunakan di Indonesia.

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas maka dapat dirumuskan masalah sebagai berikut:

1. Bagaimana membangun *dataset* yang tepat untuk digunakan di *library spacy*?
2. Bagaimana pembuatan model *dependency parser* dan *named entity recognition* bahasa Indonesia menggunakan *library spacy*?
3. Bagaimana evaluasi model *dependency parser* dan *named entity recognition*?

1.3. Batasan Masalah

Batasan masalah dari penelitian ini adalah:

1. Membuat model berdasarkan *dataset* dari *universal dependencies*.
2. Korpus yang digunakan untuk pembuatan *dataset dependency parser* dan *NER* berformat *.conllu*.

3. Format penulisan entitas pada *dataset NER* menggunakan format *offset entity*.
4. Label *dependency* yang digunakan untuk model *dependency parser* menggunakan label untuk Bahasa Inggris karena memiliki kemiripan dengan Bahasa Indonesia [4].

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah di atas maka tujuan penelitian ini adalah:

1. Membangun *dataset* yang tepat untuk digunakan di *library spacy*.
2. Membuat model *dependency parser* dan *named entity recognition* menggunakan *library spacy* bahasa Indonesia.
3. Mengevaluasi model *dependency parser* dan *named entity recognition*.

1.5. Sistematika Penulisan

Secara sistematis isi dari laporan ini disusun sebagai berikut:

BAB I : Pendahuluan

Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, dan sistematika penulisan laporan.

BAB II : Tinjauan Pustaka

Bab ini berisi penjelasan penelitian terdahulu yang sudah pernah dilakukan, serta penjelasan teori – teori dasar yang digunakan dalam penulisan laporan.

BAB III : Metodologi Penelitian

Bab ini berisi penjelasan tentang data yang akan digunakan, perangkat yang akan digunakan selama penelitian, dan langkah – langkah dalam melakukan penelitian.

BAB IV : Analisis dan Pembahasan

Bab ini berisi pembahasan dari pembuatan *dataset* untuk *dependency parser* dan *NER* juga pembuatan modelnya serta penjelasan penyatuan kedua model menjadi satu dan juga evaluasi yang dilakukan untuk kedua model tersebut. Disini juga dipaparkan analisis dari hasil pelatihan model.

BAB V : Implementasi Model Pada *Web Service*

Bab ini berisi penjelasan langkah – langkah implementasi model *dependency parser* dan *NER* pada *web service*.

BAB VI : Penutup

Bab ini berisi simpulan dari hasil penelitian yang sudah dilakukan serta saran yang mungkin dapat digunakan di masa mendatang.

