



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk mengubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

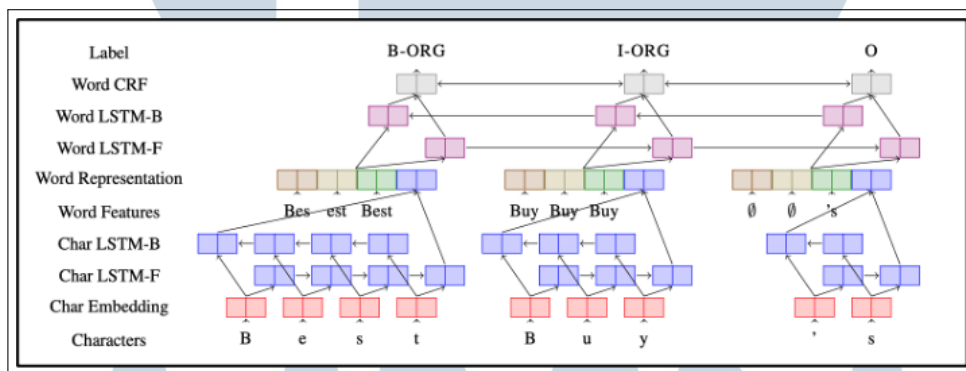
Dewasa ini, pertukaran informasi dapat berlangsung dengan sangat cepat dan mudah dengan bantuan internet. Dengan bantuan teknologi Web 2.0, setiap pengguna internet mampu membuat, menjelaskan, mengirim, dan melakukan pertukaran konten melalui sebuah halaman web [1]. Namun, hal tersebut membuat sebagian besar informasi yang beredar pada internet bersifat tidak terstruktur, contohnya adalah pada teks bebas pada surat elektronik, artikel berita, halaman web, dokumen medis, jurnal, media sosial, dan lain-lain [2]. Informasi-informasi yang terdapat pada media tersebut dipengaruhi oleh kompleksitas bahasa manusia. Hal tersebut dapat menjadi hal yang menyulitkan dalam memahami sebuah informasi. Selain itu, informasi yang tersebar di internet juga tertulis dalam berbagai macam bahasa [3]. Kosakata, struktur, dan aturan bahasa yang berbeda juga menjadi suatu tantangan dalam pemahaman dan pemrosesan sebuah informasi.

Dalam upaya mengatasi permasalahan tersebut, salah satu teknologi yang dapat digunakan adalah *Natural Language Processing* (NLP). Dengan bantuan NLP, proses ekstraksi informasi dapat berlangsung dengan lebih cepat karena menggunakan bantuan mesin atau komputer [4]. Dalam bidang NLP, terdapat beberapa tahap atau penugasan yang saling berkaitan. Menurut Sonit Singh [2], penugasan dalam NLP terbagi ke dalam tugas tingkat rendah (*low level*) dan tingkat tinggi (*high level*). Beberapa proses yang termasuk ke dalam tugas tingkat rendah adalah *POS-Tagging*, *chunking*, *parsing*, *Named Entity Recognition*, dan lain-lain. Lalu, proses yang termasuk ke dalam tugas tingkat tinggi adalah *Machine Translation*, *Information Extraction*, *Information Retrieval*, *Sentiment Analysis*, dan lain-lain. Tugas tingkat rendah menjadi dasar dari aplikasi tugas tingkat tinggi. Karena hal tersebut, performa tugas tingkat tinggi akan sangat dipengaruhi oleh keefektifan tugas tingkat rendah yang digunakan.

Salah satu tugas tingkat rendah yang biasa digunakan dalam tahap awal pada tugas tingkat tinggi seperti *question answering*, *information retrieval*, dan lain-lain adalah *Named Entity Recognition*. *Named Entity Recognition* (NER) merupakan suatu tugas untuk mengidentifikasi entitas bernama (*named entity*) dari sebuah teks, misalnya keterangan orang, lokasi, organisasi, waktu, dan lain-lain [5]. Dalam

NER, informasi-informasi tersebut diekstraksi dan diklasifikasi ke dalam kelas tertentu sebelum diproses oleh tugas berikutnya.

Saat ini, NER telah diaplikasikan dengan menggunakan berbagai macam metode. Menurut penelitian Yikas Yadav dan Steven Bethard [5], sistem NER dapat dibagi menjadi beberapa kategori berdasarkan algoritma yang digunakan, yaitu *knowledge-based systems*, *unsupervised and bootstrapped systems*, *feature-engineered supervised systems*, dan *feature-inferring neural network systems*. Hasil dari penelitian tersebut menyatakan bahwa sistem dengan model *neural network* memiliki performa yang lebih baik dibandingkan dengan model *feature-engineered*. Selain itu, dalam penelitian tersebut juga disimpulkan bahwa performa model *neural network* dapat ditingkatkan dengan melibatkan penggunaan imbuhan, karakter, dan kata sebagai fitur (*word+character+affix NN model*). Arsitektur dari *word+character+affix NN model* dapat dilihat pada Gambar 1.1.



Gambar 1.1. *Word+Character+Affix NN model*

Sumber: [5]

RNN merupakan algoritma berbasis *neural network* yang terbukti memiliki performa yang baik pada bidang NLP [6]. Namun dalam pemrosesan rangkaian data yang panjang, hasil dari RNN dapat menjadi bias karena dipengaruhi oleh *input* terbaru [7]. Oleh karena itu, arsitektur *Long Short-Term Memory* (LSTM) dikembangkan untuk mengatasi hal tersebut. LSTM menggunakan *cell state* dengan *forget gate* dan menggunakan fungsi aktivasi non-linear untuk mengatasi adanya *vanishing gradient* seperti yang terjadi pada RNN. Selain itu, dalam tugas pemrosesan data yang berurutan, perlu adanya cara untuk menganalisis data di masa lalu atau di masa depan [8]. Sama seperti RNN, LSTM juga menganalisis data dari satu arah, yaitu data sebelumnya. Maka dari itu, dikembangkan *bidirectional LSTM* (BiLSTM) yang dapat menganalisis data secara dua arah (*forward* dan *backward*). Lalu, algoritma BiLSTM disempurnakan dengan penggunaan *Conditional Random Fields*

(CRF) untuk melakukan *joint decoding* pada keseluruhan masukan yang diberikan [7].

Penggunaan algoritma BiLSTM dan CRF telah dilakukan pada beberapa penelitian NER sebelumnya, contohnya adalah penelitian Guillaume Lample, dkk. [9] dan penelitian Rrubaa Panchendrarajan dan Aravindh Amaresan [10]. Pada penelitian yang dilakukan Guillaume Lample, dkk. [9], model BiLSTM dan CRF yang dibangun memiliki akurasi yang cukup baik, yaitu sebesar 90,94%. Hasil tersebut didapat dengan mengimplementasikan *pretrained word embedding*, *character-based* model, dan *dropout* pada model yang dibangun. Lalu, pada penelitian yang dilakukan oleh Rrubaa Panchendrarajan dan Aravindh Amaresan [10], model NER dibangun menggunakan BiLSTM dan BiCRF. Hasil dari model tersebut memiliki akurasi sebesar 89,76% pada model dasarnya dan 90,84% setelah melakukan *dropout*, *POS-Tagging*, dan fitur *casing* (kapitalisasi). Penelitian tersebut menyatakan bahwa proses *dropout* meningkatkan performa model sebesar 0,635% serta pengaplikasian *POS-Tagging* dan fitur kapitalisasi meningkatkan performa model sebesar 0.35%. Maka dari itu, penggunaan beberapa fitur tersebut akan dipertimbangkan dalam penelitian ini.

Berdasarkan penelitian yang sudah dilakukan sebelumnya, penelitian ini akan dilakukan untuk melakukan tugas *Named Entity Recognition*, khususnya untuk menghadapi permasalahan ekstraksi informasi dalam beberapa bahasa. Pada penelitian ini akan dibangun sebuah model NER berbasis algoritma BiLSTM-CRF. Algoritma BiLSTM-CRF dipilih karena memiliki performa yang cukup baik pada penelitian-penelitian sebelumnya. Selain itu, penelitian ini akan menguji dan membandingkan performa model dengan atau tanpa penggunaan fitur *POS-Tag* dan pelatihan dengan *bilingual dataset*. Penggunaan fitur *POS-Tag* dan pelatihan dengan *bilingual dataset* diharapkan mampu meningkatkan performa model dalam melakukan klasifikasi NER.

1.2. Rumusan Masalah

Berdasarkan latar belakang penelitian, ada beberapa permasalahan yang dapat ditarik, antara lain:

1. Bagaimana cara membangun model NER menggunakan algoritma BiLSTM-CRF?
2. Bagaimana performa model NER berbasis algoritma BiLSTM-CRF dalam

melakukan klasifikasi entitas bernama dari suatu kata dengan atau tanpa aplikasi POS-Tag dan *bilingual dataset*?

1.3. Batasan Permasalahan

Pada penelitian ini, terdapat beberapa batasan masalah yang diterapkan, antara lain:

1. Model yang dibangun akan dilatih menggunakan 2 dataset dengan 2 bahasa yang berbeda, yaitu bahasa Indonesia dan bahasa Inggris.
2. Dataset Bahasa Indonesia yang digunakan pada penelitian ini adalah dataset SINGGALANG [11]. Dataset tersebut terdiri dari 48.597 kalimat dengan 3 jenis entitas.
3. Dataset Bahasa Inggris yang digunakan pada penelitian ini adalah dataset CoNLL-2003 (*English*) [12]. Dataset tersebut terdiri dari 20.744 kalimat dengan 4 jenis entitas berbasis *BIO-Format*. Dataset tersebut juga menyimpan atribut POS tag dan *chunk tag* dari setiap kata.
4. Entitas pada model *Named Entity Recognition* yang dibangun dibatasi menjadi 3 kelas, sesuai dengan jumlah entitas paling sedikit dari kedua dataset. Entitas yang akan digunakan yaitu *Person (PER)*, *Location (LOC)*, *Organization (ORG)*. Ketiga entitas tersebut akan menggunakan bentuk *BIO-Format*. Selain itu, digunakan *tag Others (O)* untuk mewakili entitas lain diluar 3 entitas sebelumnya.
5. *POS-Tagging* hanya dilakukan pada dataset Bahasa Indonesia karena dataset Bahasa Inggris sudah memiliki atribut POS-Tag. *POS-Tagging* dilakukan menggunakan *pretrained POS-Tagger model*.

1.4. Tujuan Penelitian

Pada penelitian ini terdapat beberapa tujuan yang hendak dicapai, yaitu:

1. Membangun model NER dengan algoritma BiLSTM-CRF.
2. Mengukur tingkat akurasi model NER yang dibuat dengan atau tanpa aplikasi POS-Tag dan *bilingual dataset*.

1.5. Manfaat Penelitian

Adapun beberapa manfaat yang dapat diperoleh dari penelitian ini, antara lain:

1. Membangun model NER berbasis algoritma BiLSTM-CRF yang dapat digunakan untuk melakukan klasifikasi entitas bernama.
2. Menjadi referensi untuk penelitian di bidang NLP atau NER berikutnya.

1.6. Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN

Bab 1 berisikan latar belakang penelitian, rumusan dan batasan masalah penelitian, tujuan dan manfaat penelitian, dan sistematika penulisan laporan.

- Bab 2 LANDASAN TEORI

Bab 2 membahas literatur-literatur yang mendasar dalam penelitian ini, antara lain teori tentang cara merancang dan membangun model NER, algoritma BiLSTM, algoritma CRF, pemrosesan POS-*Tagging*, perhitungan metrik evaluasi, dan informasi-informasi lainnya.

- Bab 3 METODOLOGI PENELITIAN

Bab 3 mendeskripsikan rancangan model yang akan digunakan untuk merepresentasikan model NER yang akan dibangun. Rancangan model yang akan dibangun digambarkan ke dalam bentuk *flowchart*.

- Bab 4 HASIL DAN DISKUSI

Bab 4 menjelaskan tentang implementasi sistem yang dibangun dalam bentuk *source code*. Selain itu, terdapat skenario pengujian, hasil pengujian, dan evaluasi dari hasil yang diperoleh.

- Bab 5 KESIMPULAN DAN SARAN

Bab 5 berisikan kesimpulan dari hasil penelitian serta saran yang dapat dilakukan untuk penelitian lebih lanjut.