



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 I

PENDAHULUAN

1.1 Latar Belakang

Bahasa Indonesia tergolong dalam bahasa aglutinasi (Ekananda, 2019). Bahasa aglutinasi adalah bahasa yang kata dasarnya banyak mengalami pembentukan kata seperti pada proses afiksasi, reduplikasi, dan komposisi. Pembentukan kata dilakukan agar kata bisa dipergunakan dalam kalimat (Chaer, 2008). Proses pembentukan kata merupakan bidang kajian morfologi dalam ilmu linguistik. Melihat bahasa Indonesia yang adalah bahasa aglutinasi, bahasa Indonesia bisa dibilang kaya akan aspek morfologi atau Morphologically Rich Languages (MRL). MRL mempunyai beberapa faktor linguistik yang melekat antara lain mempunyai bentuk kata yang beragam, penempatan kata yang lebih bebas, dan morfologi kata yang menyiratkan hubungan sintaktis. Pengertian hubungan sintaktis adalah makna gramatik yang timbul ketika melakukan penggabungan kata menjadi satuan bahasa yang lebih besar seperti frasa atau klausa dalam kalimat (Tsarfaty, dkk., 2010).

Penguraian dependensi atau *dependency parsing* merupakan salah satu tugas di dalam *Natural Language Processing* (NLP) yang melakukan analisis sintaksis kalimat (Sciforce, 2018). Tugas penguraian dependensi erat hubungannya dengan morfologi kata, sehingga keduanya tidak terpisahkan. Tujuan penguraian dependensi adalah menghasilkan pohon dependensi. Pohon dependensi memberitahukan hubungan

antara satu kata dan kata lainnya secara fungsional (Kübler, dkk., 2009). Pada analisis sintaksis kata dapat berfungsi sebagai objek, subjek, predikat, dan lain-lain.

Metode pengurai dependensi umumnya terbagi dua yaitu *transition-based* dan *graph-based*. Metode *transition-based* menggunakan suatu sistem transisi yang terdiri dari konfigurasi dan transisi. Proses penguraian dependensi pada metode *transition-based* melibatkan penerapan transisi terhadap suatu konfigurasi non terminal agar dapat mencapai konfigurasi terminal. Pengurai dependensi memberikan skor terhadap segala kemungkinan transisi. Transisi dengan skor tertinggi dipilih untuk dijalankan pada sistem transisi. Karena pengurai dependensi selalu memilih transisi yang tertinggi dan mengabaikan transisi lain yang mungkin hasilnya lebih baik maka penguraian ini sifatnya *greedy*. Pohon dependensi didapatkan dari konfigurasi terminal pada suatu sistem transisi. Sedangkan metode *graph-based* mendefinisikan ruang yang berisikan pohon-pohon dependensi yang valid. Penguraian pada metode *graph-based* menghitung skor sub-sub pohon dari suatu pohon dependensi dan kemudian dijumlahkan. Pohon dependensi yang dipilih adalah pohon dengan penjumlahan skor sub-sub pohonnya tertinggi. Batasan umum yang ada pada pengurai dependensi salah satunya adalah kalimat harus bersifat projektif. Yang dimaksud dengan projektif adalah hubungan antara kata penjelas dengan kata yang diperjelas tidak boleh diselah oleh kata lainnya (Kübler, dkk., 2009).

Program pengurai dependensi kebanyakan dikembangkan untuk mengurai kalimat bahasa Inggris seperti yang dibuat oleh Chen dan Manning (2014). Bahasa Inggris merupakan bahasa analitik yang artinya ketatabahasaannya lebih ditekankan

pada sintak sehingga tidak termasuk dalam MRL. Bila pengurai dependensi bahasa Inggris diadaptasi untuk mengurai kalimat MRL hasilnya tidak sebaik seperti saat mengurai kalimat bahasa Inggris (Tsarfaty, dkk., 2010). Tsarfaty, dkk. (2013) menambahkan lagi bahwa analisis morfologi diperlukan pada kalimat MRL sebagai upaya peningkatan performa. Agar hasil pengurai dependensi untuk bahasa Indonesia bisa lebih maksimal aspek morfologi harus diikutsertakan dalam penguraian dependensi seperti yang diutarakan oleh Tsarfaty, dkk. (2013).

Pengurai dependensi yang dikembangkan Chen dan Manning (2014) tidak mengikutsertakan analisis morfologi secara khusus. Kalimat yang diurai oleh pengurai dependensi direpresentasikan menggunakan vektor kata dari *word2vec* (Mikolov, dkk., 2013) atau GloVe (Pennington, dkk., 2014). Kata yang direpresentasikan oleh *word2vec* atau GloVe menyimpan informasi sintaksis dan semantik di tingkat kata. Untuk bisa mengetahui morfologi, kata perlu direpresentasikan sampai ke tingkat karakter (Smith, dkk., 2018; Vania, dkk., 2018). Bentuk representasi kata lain yaitu ELMo (Peters, dkk., 2018) dapat menghasilkan vektor kata sampai ke tingkat karakter. ELMo merupakan *language model* yang terdiri atas dua layer *Bidirectional Long Short-Term Memory* (BiLSTM) dan sebuah *Convolutional Neural Network* (CNN). Setiap layer pada ELMo menghasilkan representasi kata. Representasi kata pada CNN layer dikomposisikan dari vektor-vektor karakter yang menjadi pembentuk kata. Pengkomposisian ini mirip dengan proses pembentukan kata pada proses morfologi. Dengan ditambahkannya aspek morfologi dengan menggunakan representasi kata

tingkat karakter milik ELMo, maka diharapkan penelitian ini dapat meningkatkan performa pengurai dependensi bahasa Indonesia

1.2 Rumusan Masalah

Berdasarkan latar belakang yang tertuang di atas maka rumusan masalah dari penelitian ini adalah sebagai berikut.

- a. Bagaimana mengimplementasikan representasi kata tingkat karakter dengan menggunakan ELMo pada pengurai dependensi bahasa Indonesia ?
- b. Berapa *Unlabeled Attachment Score* (UAS) dan *Labeled Attachment Score* (LAS) dari pengurai dependensi bahasa Indonesia yang mengimplementasikan representasi kata tingkat karakter dengan menggunakan ELMo .

1.3 Batasan Masalah

Pengurai dependensi yang dibangun mempunyai beberapa batasan antara lain.

- a. Metode penguraian menggunakan *transition-based dependency parsing*.
- b. Representasi kata yang digunakan dari ELMo adalah representasi kata tingkat karakter di layer CNN.
- c. Pengurai dependensi menggunakan model Feed-Forward Network (FNN).
- d. Penguraian dilakukan secara *greedy* dan hanya mengurai kalimat yang pohon dependensinya bersifat projektif.

1.4 Tujuan Penelitian

Penelitian ini dilakukan dengan tujuan sebagai berikut

- a. Mengimplementasikan representasi kata tingkat karakter dengan menggunakan ELMo pada pengurai dependensi bahasa Indonesia

- b. Mengetahui UAS dan LAS dari pengurai dependensi bahasa Indonesia yang mengimplementasikan representasi kata tingkat karakter dengan menggunakan ELMo

1.5 Manfaat Penelitian

Hasil penelitian ini diharapkan bisa menjadi bahan pertimbangan dalam membangun pengurai dependensi bahasa Indonesia pada penelitian selanjutnya, dan menyediakan pengurai dependensi bahasa Indonesia yang sudah dilatih agar bisa digunakan oleh publik atau tugas hilir NLP yang membutuhkan analisis sintaksis sebagai fitur masukan.

1.6 Sistematika Penulisan Laporan Penelitian

Sistematika penulisan yang digunakan dalam laporan ini adalah sebagai berikut

- a. Bab I (Pendahuluan)

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

- b. Bab II (Landasan Teori)

Bab ini berisi tentang seluruh landasan teori yang digunakan dalam penelitian ini

- c. Bab III (Metode dan Perancangan Sistem)

Bab ini berisi tahapan perancangan yang berisikan diagram-diagram, dan *mockup* tampilan program

- d. Bab IV (Implementasi dan Pengujian)

Bab ini berisi aktualisasi rancangan ke penulisan program ke bentuk kode dan disertai dengan uji coba program.

e. Bab V (Simpulan dan Saran)

Bab ini berisi simpulan dari hasil penelitian dan saran untuk penelitian selanjutnya.