

Identifikasi Kata Majemuk Bahasa Indonesia

Fikri Haykal¹, Arie Ardiyanti Suryani², Sri Widowati³

^{1,2,3}Fakultas Informatika, Universitas Telkom, Bandung

¹fikrihaykal@students.telkomuniversity.ac.id, ²ardiyanti@telkomuniversity.ac.id,

³sriwidowati@telkomuniversity.ac.id

1. Pendahuluan

Latar Belakang

Multi-word expression merupakan sebuah kata yang menghasilkan makna baru yang terdiri dari setidaknya dua kata dasar yang digabungkan [7]. Dalam bahasa Indonesia contoh dari *Multi-word expression* adalah kata majemuk karena, gabungan morfem dasar yang seluruhnya berstatus sebagai kata yang mempunyai pola bunyi, gramatikal dan semantis yang khusus menurut kaidah bahasa yang bersangkutan. Pola khusus tersebut membedakannya dari gabungan morfem dasar yang bukan kata majemuk, atau dengan kata lain kata majemuk merupakan hasil proses perpaduan dua unsur kata yang mengandung satu makna baru [4, 5]. Pada tugas akhir ini dilakukan pembangunan mesin identifikasi kata majemuk dengan acuan aplikasi POS *tagger* yang dibangun oleh Universitas Indonesia, karena, pada POS *tagger* tersebut jika menggunakan *multi-word expression tokenizer* dengan CRF *tagger* akurasi yang didapat sebesar 70% sedangkan, jika tidak menggunakan *multiword expressions tokenizer* tersebut akurasi POS *tagger* sebesar 79% [1]. Oleh karena itu pada tugas akhir ini dilakukan pembangunan mesin identifikasi kata majemuk berbahasa Indonesia dengan menggunakan tiga POS *tagger* dengan metode yang berbeda, yaitu, CRF *tagger*, Bigram *tagger* dan *Classifier Based tagger* untuk mengetahui apakah POS *tagger* dengan tiga metode yang berbeda tersebut akan memengaruhi hasil akhir dari mesin identifikasi kata majemuk berbahasa Indonesia yang dibangun ini dan akurasi hasil dapat mencapai $\geq 80\%$ atau tidak, dan juga dengan harapan dapat diimplementasikan pada POS *Tagger* bahasa Indonesia yang ada.

Topik dan Batasannya

Topik permasalahan pada tugas akhir ini adalah menurunnya akurasi ketepatan pada POS *tagger* yang dibangun oleh Universitas Indonesia, jika menggunakan MWE *tokenizer* dengan metode CRF *tagger* akurasi ketepatan bernilai 70%, sedangkan jika tidak menggunakan MWE *tokenizer* tersebut akurasi ketepatan bernilai 79%. Hal tersebut berdampak pada hasil yang didapatkan akan menjadi kurang baik, karena, ada beberapa kata yang seharusnya merupakan kata majemuk tetapi mesin tidak dapat mengidentifikasi kata majemuk tersebut. Oleh karena itu pada tugas akhir ini akan dilakukan pembangunan mesin identifikasi kata majemuk dengan harapan akurasi hasil sebesar $\geq 80\%$, adapun batasannya, mesin identifikasi kata majemuk ini hanya berbahasa Indonesia, hanya untuk kata majemuk *non-senyawa*, evaluasi dilakukan secara manual berlandaskan pola dan karakteristik kata majemuk yang sudah didefinisikan oleh ahli bahasa Indonesia, dan mengabaikan abiguitas kata majemuk tersebut.

Tujuan

Tujuan tugas akhir ini adalah mesin yang di bangun dapat menghasilkan akurasi hasil sebesar $\geq 80\%$ dengan harapan dapat digunakan untuk diimplementasikan pada POS *Tagger* berbahasa Indonesia yang ada.

Organisasi Tulisan

Pada jurnal ini dibagi menjadi 4 bagian, yaitu;

1. Studi Terkait
Bagian ini berisi teori/studi/literatur yang mendukung (terkait erat) dengan topik TA yang dikerjakan. Bagian ini bisa bernama Tinjauan Pustaka atau Landasan Teori. Dalam bahasa Inggris disebut sebagai Related Work atau Literature Review.
2. Sistem yang Dibangun
Pada bagian ini dijelaskan rancangan dan sistem yang dihasilkan
3. Evaluasi
Pada bagian ini berisi dua sub-bagian, yaitu Hasil Pengujian dan Analisis Hasil Pengujian. Pengujian dan analisis yang dilakukan selaras dengan tujuan TA sebagaimana dinyatakan dalam Pendahuluan.
4. Kesimpulan
Pada bagian ini berisi kesimpulan dan saran (*future works*) mengenai hasil dan apa yang bisa dilakukan berikutnya pada hasil pengerjaan ini.