Development of Annotation Guidelines, Treebanks, and a Tree Rotation Method that Conform to Universal Dependencies v2 for Indonesian Dependency Parsing = Pengembangan Petunjuk Anotasi, Treebank dan Metode Rotasi Tree yang Mengacu ke Universal Dependencies v2 untuk Dependency Parsing Bahasa Indonesia

Ika Alfina, author

Abstrak

Pada penelitian ini, kami ingin mengatasi masalah langkanya dataset untuk peneli- tian di bidang syntactic parsing untuk Bahasa Indonesia, terutama kurang tersedi- anya dependency treebank berbahasa Indonesia dalam kualitas yang baik. Adapun tujuan dari penelitian ada tiga: 1) mengusulkan petunjuk cara menganotasi depen- dency trebank untuk Bahasa Indonesia yang mengacu kepada aturan anotasi UD v2, 2) membangun dependency treebank yang dianotasi secara manual agar bisa berperan sebagai gold standard, 3) membangun sebuah dependency treebank de- ngan mengkonversi secara otomatis sebuah constituency treebank menjadi sebuah dependency treebank.

Kami sudah membuat panduan anotasi untuk membangun dependency treebank untuk Bahasa Indonesia yang mengacu kepada aturan UD v2. Pedoman tersebut mencakup aturan tokenisasi/segmentasi kata, pelabelan kelas kata (POS tagging), analisis fitur morfologi, dan anotasi hubungan dependency antar kata. Kami men- gusulkan bagaimana memproses klitika, kata ulang, dan singkatan pada tahap to- kenisasi/segmentasi kata. Pada tahapan penentuan kelas kata, kami mengusulkan pemetaan dari daftar kata dalam Bahasa Indonesia ke 17 kelas kata yang didefin- isikan oleh UD v2. Untuk anotasi fitur morfologi, kami telah memilih 14 dari 24 fitur morfologi UD v2 yang dinilai sesuai dengan aturan Bahasa Indonesia, berikut dengan 27 buah label feature-value yang bersesuaian dengan fitur morfologi terkait. Untuk anotasi hubungan dependency antarkata, kami mengusulkan penggunakan 14 buah label yang bersifat language-specific untuk menganotasi struktur sintaks yang khusus terdapat pada Bahasa Indonesia.

Sebuah dependency treebank berbahasa Indonesia yang bisa digunakan sebagai gold standard sudah berhasil dibangun. Treebank ini dibuat dengan merevisi se- cara manual sebuah dependency treebank yang sudah ada. Revisi dilakukan dalam dua fase. Pada fase pertama dilakukan koreksi terhadap tokenisasi/segmentasi kata, pelabelan kelas kata, dan anotasi terhadap hubungan dependency antarkata. Pada fase kedua, selain dilakukan sedikit koreksi untuk perbaikan pada tahap satu, di- tambahkan juga informasi kata dasar (lemma) dan fitur morfologi. Evaluasi ter- hadap kualitas treebank yang baru dilakukan dengan membangun model depen- dency parser menggunakan UDPipe. Hasil pengujian menunjukkan bahwa kami berhasil meningkatkan kualitas treebank, yang ditunjukkan dengan naiknya UAS sebanyak 9% dan LAS sebanyak 14%.

Terkait tujuan penelitian ketiga, kami juga sudah membangun sebuah treebank baru dengan mengkonversi secara otomatis sebuah constituency treebank ke dependency treebank. Pada proyek ini, kami mengusulkan sebuah metode rotasi tree yang bertu- juan mengubah dependency tree awal yang dihasilkan oleh alat NLP untuk Ba- hasa Inggris bernama Stanford UD converter sedemikan agar head-directionality dari frase kata benda yang dihasilkan sesuai dengan aturan Bahasa Indonesia yang umumnya bersifat head-initial. Kami menamakan algoritma yang dihasilkan seba- gai algoritma headSwap dan algoritma compound. Hasil

percobaan menunjukkan bahwa metode rotasi tree yang diusulkan berhasil meningkatkan performa UAS sebanyak 32.5%.

.....In this dissertation, we address the lack of resources for Indonesian syntactic parsing research, especially the need for better quality Indonesian dependency treebanks. This work has three objectives: 1) to propose annotation guidelines for Indonesian dependency treebank that conform to UD v2 annotation guidelines, 2) to build a gold standard dependency treebank, 3) to build a silver standard dependency tree- bank by converting an existing Indonesian constituency treebank automatically to a dependency treebank.

We have proposed a set of annotation guidelines for Indonesian dependency tree- bank that conform to UD v2. The guidelines cover tokenization/word segmenta- tion, POS tagging, morphological features analysis, and dependency annotation. We proposed how to handle Indonesian clitics/multiword tokens, reduplication, and abbreviation for word segmentation. For POS tagging, we presented the mapping from UD v2 guidelines to the Indonesian lexicon. For morphological features, we proposed the use of 14 of 24 UD v2 morphological features along with 27 UD v2 feature-value tags for Indonesian grammar. Finally, we proposed using 14 language- specific relations to annotate the particular structures in Indonesian grammar for dependency annotation.

A gold standard Indonesian dependency treebank also has been built based on our proposed annotation guidelines. The gold standard was constructed by manually revised an existing Indonesian dependency treebank. The revision project consists of two phases. Major revision on word segmentation, POS tagging, and dependency relation annotation was conducted in the first phase. In the second phase, we added the lemma information and morphological features. Finally, we evaluated the qual- ity of the revised treebank by building a dependency parser using UDPipe. The experiment results show that we successfully improved the quality of the original treebank with a margin of 9% for UAS and 14% for LAS.

Finally, we built a silver standard treebank by automatically converting an Indone- sian constituency treebank to a dependency treebank. In this work, we proposed a method to improve the output of an English NLP tool named Stanford UD con- verter. We transformed the output so that it conforms to the head-directionality rule for noun phrases in Indonesian. We called the proposed tree rotation algorithm the headSwap method and the rule for noun phrases as the compound rule. The evaluation shows that our proposed method improved the UAS with a margin of 32.5%.