Pengembangan tokenizer dan morphological analyzer universal untuk Bahasa Indonesia menggunakan Finite-State Transducer = Building universal tokenizer and morphological analyzer for Indonesian Language with Finite-State Transducer

Muhammad Yudistira Hanifmuti, author

Abstrak

Morphological analyzer merupakan sebuah alat yang digunakan untuk melihat bagaimana proses pembentukan kata, menentukan kata dasar pembentuk, dan menge- tahui informasi linguistik yang terkandung pada suatu kata. Universal Dependencies (UD) merupakan sebuah framework acuan yang digunakan pada proses anotasi morfologi untuk berbagai bahasa. Sayangnya, belum ditemukan morphological analyzer untuk bahasa Indonesia yang menerapkan pedoman UD ini. Penelitian ini mengembangkan morphological analyzer untuk bahasa Indonesia yang diberi nama Aksara. Aksara dibangun menggunakan finite state compiler bernama Foma yang digunakan pada Mor- phind, morphological analyzer pada penelitian sebelumnya. Foma dapat memodelkan aturan-aturan pembentukan kata dalam bentuk finite state transducer. Pada Aksara juga dikembangkan tokenizer yang hasilnya menyesuaikan dengan hasil tokenisasi pada treebank UD. Implementasi Aksara menerapkan pedoman UD versi terbaru yaitu UDv2. Pengujian Aksara dilakukan dengan membandingkan performa Aksara dengan Morhpind. Hasil pengujian menunjukkan bahwa komponen tokenizer Aksara berhasil memiliki akurasi tokenisasi sebesar 96.60%, meningkat 23.89% dari akurasi tokenisasi oleh Mor- phind. Evaluasi POS tagging Aksara juga berhasil melewati hasil pemetaan Morphind dengan akurasi F1-score sebesar 87%, dengan kenaikan relatif sebesar 18% dari baseline.

.....Morphological analyzer is a tool used to do an analysis on word formation process, to identify the lemma for each word, and to do an analysis on the linguistic information. Universal Dependencies (UD) is a framework commonly used in morphological annota- tion process. Unfortunately, there is not a single Indonesian morphological analyzer that applies UDv2. This research is a development of morphological analyzer for Indonesian language named Aksara. Aksara was build using finite state compiler named Foma, which was used in Morphind, the previous research on Indonesian morphological analyzer. Foma can model the rules of word formation which is represented in the form of finite state transducer. This research also develops a tokenizer which its results are adjusted to the tokenization example on UD treebank. The Aksara implementation applies the latest UD guidelines, UDv2. Testing of Aksara is done by comparing the performance of Aksara with Morphind. The test results show that the tokenizer component of Aksara managed to have a tokenization accuracy of 96.60%, an increase of 23.89% from the accuracy of tokenization by Morphind. Evaluation of POS tagging with Aksara also managed to pass Morphind with an accuracy of F1-score of 87%, with a relative increase of 18% from the accuracy of Morphind.