Penanganan Noisy Text untuk Meningkatkan Akurasi Lemmatisasi dan POS Tagging untuk Bahasa Indonesia Informal = Handling Noisy Text to Improve Lemmatization and POS Tagging Accuracy for Informal Indonesian Text

Erica Harlin, author

Abstrak

Aksara adalah sebuah NLP tool yang menuruti Universal Dependencies (UD) v2. Penelitian terakhir terkait pemrosesan bahasa informal pada Aksara adalah v1.2 yang berfokus pada kemampuan Aksara untuk memproses kata-kata dasar informal dan kata-kata dengan afiksasi informal. Penelitian ini bertujuan untuk mengembangkan kemampuan Aksara dalam memproses noisy text. Dalam penelitian ini, terdapat 5 metode yang dipertimbangkan untuk menormalisasikan noisy text, yaitu: Levenshtein distance, Damerau-Levenshtein distance, perbandingan subsequence, longest common subsequence (LCS), dan SymSpell. Untuk menentukan metode mana yang paling cocok, kami membangun dataset sintetis berukuran 20.000 kata, lalu mengukur dan membandingkan performa metode yang satu dengan yang lain dalam menormalisasikan dataset sintetis tersebut. Metode yang akhirnya dipilih adalah SymSpell karena metode ini yang menghasilkan akurasi yang paling tinggi. Versi Aksara yang dihasilkan oleh penelitian ini adalah Aksara v1.4 (Aksara baru). Untuk mengevaluasi Aksara baru, dipakai gold standard yang terdiri dari 152 kalimat dan 1786 token. Hasil evaluasi menunjukkan lemmatizer Aksara baru memiliki akurasi senilai 90.99% dan 91.66% untuk kasus case-sensitive dan case-insensitive. Untuk POS tagger, Aksara baru memiliki akurasi senilai 83%, recall senilai 83%, dan F1 score senilai 83%.

.....

Aksara is an Indonesian NLP tool that conforms to Universal Dependencies (UD) v2. The latest work on Aksara pertaining to its informal language processing ability is Aksara v1.2, which is focused on Aksara’s ability to process informal root words and words with informal affixation. This work aims to enable Aksara to process noisy texts. In this research, there are 5 methods considered for normalizing noisy texts: Levenshtein distance, Damerau-Levenshtein distance, subsequence comparison, longest common subsequence (LCS), and SymSpell. To determine which method is best suited for this purpose, we built a synthetic dataset of 20,000 words, then measured and compared each method’s performance in normalizing the synthetic data. The chosen method is SymSpell as it yields the highest accuracy. This chosen method along with a context dictionary will be integrated into Aksara as a text normalizer. To evaluate new Aksara’s performance, a gold standard consisting of 152 sentences and 1786 tokens is used. The evaluation result shows that the new Aksara’s lemmatizer has an accuracy of 90.99% and 91.61% for case-sensitive and case-insensitive cases. For POS tagger, the new Aksara has an accuracy of 83%, a recall of 83%, and an F1 score of 83%.