Deskripsi Tugas II Pemrosesan Bahasa Alami

Topik Terkait: POSTagging

Judul Tugas: Membangun POSTagger Bahasa Indonesia

Batas Pengumpulan: 6 Oktober 2018 (23:59)

Deskripsi:

Buatlah POSTagger Bahasa Indonesia yang dapat menerima masukan berupa kalimat, dan akan menghasilkan label POSTag untuk tiap kata dalam kalimat masukan tersebut.

Untuk membangun sebuah POSTagger, terdapat beberapa pendekatan yang dapat diaplikasikan. Berikut adalah pendekatan/metode yang wajib diaplikasikan dalam POSTagger yang Anda buat:

1. Metode baseline (berdasar frekuensi tag sebuah kata yang paling sering ditemui pada korpus latih).
2. Metode berbasis statistika, persoalan POSTagging diselesaikan sebagai sebuah persoalan klasifikasi. (contoh yang diberikan dengan metode *decision tree*). Anda boleh melakukan eksperimen dengan fitur dan atau metode klasifikasi yang lain.
3. Metode berbasis pemodelan sekuens, HMM-Viterbi.

Dataset latih yang digunakan adalah korpus POSTag yang dapat diunduh dari http://bahasa.cs.ui.ac.id/postag/corpus , cukup ambil **1000 kalimat pertama sebagai data latih**, dan ambil **20 kalimat berikutnya sebagai data uji**. Lakukan pengujian terhadap 3 POSTagger yang Anda buat dengan pendekatan yang berbeda-beda tersebut, dengan mengukur nilai akurasi.

Deliverables:

1. Softcopy program beserta komentar yang mudah dimengerti
2. Petunjuk cara eksekusi program
3. Laporan singkat yang berisi:
   a. Keterangan singkat metode/pendekatan yang diterapkan, terutama jika mengusulkan gabungan pendekatan atau menerapkan pendekatan yang lain.
   b. Analisis terhadap akurasi POSTagger dari hasil pengujian.

Komponen penilaian:

1. Kebenaran, kelengkapan dan kejelasan program (nilai maksimum: 70)
2. Laporan (nilai maksimum: 30)
3. Bonus: penerapan gabungan beberapa pendekatan (contoh: mendefinisikan langkah pemrosesan POSTagging, sebagai contoh seperti yang diterapkan pada paper Rule-Based Indonesian POSTagger), menerapkan metode POSTagging yang lain (misal: penggunaan aturan yang didefiisikan secara manual, penggunaan kamus untuk memproses closed-class POSTag, dll) (nilai maksimum: 20)