

Nama: Raden Rizky Falih

NIM: 1301154211

LAPORAN

TUGAS 2 PEMROSESAN BAHASA ALAMI

1. Metode

Metode yang digunakan terdapat 3 buah metode, yaitu baseline, statistical dan viterbi. Untuk lebih jelasnya akan dijelaskan pada sub bab berikut:

- Baseline Model

Baseline model merupakan suatu pembentukan model untuk menentukan POST Tag dari sebuah kata yang akan diprediksi. Dalam metode ini, penentuan prediksi dilakukan berdasarkan banyaknya tag yang muncul untuk kata tersebut. Misal kata 'saya' memiliki 2 buah tag, yaitu 'A' dan 'B'. Kemudian kemunculan kata 'saya' dengan tag 'B' lebih banyak daripada tag 'A'. Maka dapat disimpulkan kata 'saya' memiliki prediksi tag berupa 'B'.

- Statistical Model

Statistical model merupakan suatu pembentukan model untuk menentukan POST Tag dari sebuah kata yang akan diprediksi berbasis dengan data statistik masing-masing. Dalam metode ini, penentuan prediksi dilakukan dengan metode klasifikasi (metode yang ada dalam machine learning). Dalam hal ini penulis menggunakan metode Decision Tree untuk membangun model statistical tersebut.

- Viterbi (Sekuens)

Sama halnya seperti 2 model sebelumnya, tujuan dari penggunaan viterbi adalah untuk membangun suatu model agar bisa memprediksi POST Tag dari sebuah kata. Dalam pembentukan modelnya, viterbi membutuhkan nilai probabilitas antar setiap kata dengan tagnya dan nilai probabilitas antar setiap tag. Dari kebutuhan nilai tersebut dapat dianalisa bahwa viterbi adalah metode yang dibentuk berdasarkan probabilitas. Untuk menentukan keputusannya, viterbi mengambil kemungkinan yang memiliki nilai probabilitas yang sangat besar.

2. Analisi:

Dari percobaan yang dilakukan didapat nilai akurasi **Baseline model 82%**, nilai akurasi **Statistical Model (Decision Tree) 93%** dan nilai **akurasi Viterbi 25%**. Dapat dilihat bahwa statistical Model memiliki nilai probabilitas paling tinggi, hal itu dikarenakan pada Statistical Model penulis menggunakan fitur imbuhan, kata sebelum dan kata sesudah (fitur suffix) sehingga untuk kompleksitas model yang dibentuk lebih baik dari 2 model tersebut. Sedangkan viterbi memiliki akurasi paling kecil sebesar 25%, hal itu dikarenakan masih ada kesalahan algoritma pada model viterbi sehingga accuracy yang dihasilkan tidak begitu bagus. Akan tetapi menurut hipotesa saya, apabila algoritma berjalan dengan baik seharusnya nilai akurasi viterbi lebih baik daripada baseline dan tidak lebih baik dari statistical. Hal itu disebabkan viterbi memerhatikan tata urutan suatu kata dan keterhubungan setiap post tag, sehingga ia memiliki nilai probabilitas masing-masing di setiap sekuensnya. Sedangkan baseline tidak memiliki nilai probabilitas tersebut, hanya memerhatikan seberapa banyak kata tersebut dengan nilai tag tertentu.

Kesimpulan:

Dari analisa/hipotesa tersebut dapat disimpulkan bahwa Statistical model lebih baik dari metode lainnya. Sedangkan baseline model memiliki nilai akurasi terendah (apabila hipotesa tentang viterbi benar) karena hanya memerhatikan banyaknya kemunculan tag di suatu kata.