Nama : Fero Resyanto

Nim : 1301154318

LAPORAN TUGAS 2

Membangun POSTagger Bahasa Indonesia

1. Dataset

Data set yang digunakan untuk membagun POSTTagger bahasa Indonesia ini diambil dari korpus posttag yang diunduh dari link: <http://bahasa.cs.ui.ac.id/postag/corpus> , data yang diunduh sebanyak 10.000 kalimat. Dari 10.000 kalimat tersebut diambil 1000 kalimat pertama untuk dijadikan data *training*  dan 20 kalimat selanjutnya diambil untuk dijadikan data *test*. Data *training* disimpan pada file *data\_training.txt* dan data *test* disimpan pada file *data\_tes.txt*. data *training* dan data *testing* digunakan pada seluruh metode posttag yang dilakukan (baseline,klasifikasi dan viterbi).

1. Metode

Metode pertama yang digunakan yaitu metode **baseline**, dimana metode ini akan memberi tag pada suatu kata dengan melihat tag kata tersebut pada data *training*. Pada metode baseline, jika terdapat kata yang tidak ada dalam data *training* maka kata tersebut akan diberi tag ‘NN’ oleh program. Metode kedua yang digunakan yaitu metode **klasifikasi** dengan **Decision Tree**. Metode ini akan memberi tag pada data *test* dengan cara melakukan klasifikasi berdasarkan *learning* dari data *training* Decision Tree. Metode ketiga yaitu menggunakan **HMM** (hidden markov model) algoritma **Viterbi** untuk mencari tag pada data *test*. Algoritma Viterbi menghitung nilai emmision dan transition probabilty yang digunakan untuk mencari tag pada kata. Jika terdapat kata yang tidak ada dalam data *training*, maka program akan memberi kata tersebut tag ‘NN’. Tag ‘NN’ diberikan karena tag tersebut merupakan tag dengan jumlah terbesar dari seluruh tag yang ada.

1. Analisis dan kesimpulan

Hasil pengujian posttager dengan metode baseline mendapatkan nilai akurasi 87,975 %. Sedangkan untuk pengujian posttager dengan metode klasifikasi decision tree mendapatkan nilai akurasi 94,789% . dan metode terkahir yaitu model HMM mendapatkan nilai akurasi sebesar 87,374 %. Menurut pemahaman saya, model klasifikasi memiliki nilai akurasi tertinggi karena terdapat proses learning pada metode decision tree sehingga pemberian tag lebih akurat terutama pada kata-kata yang belum dikenal. Karena metode klasifikasi decision tree dapat mengklasifikan kata yang belum ada pada tree yang terbentuk. Sedangkan model baseline dan HMM viterbi memiliki nilai akurasi yang hampir sama yaitu sekitar 87%. Hal ini dikarenakan jika terdapat kata baru maka sistem akan secara otomatis memberi tag ‘NN’ tanpa adanya proses learning yang lebih baik. Walaupun pada model viterbi melihat tag kata sebelumnya (transition probabilty) namun tidak memberi hasil yang lebih baik dari model baseline maupun model klasifikasi. Untuk model baseline sendiri memiliki kekurangan jika data *training* yang dimiliki sedikit. Menurut saya, Semakin banyak data training maka akan meningkatkan nilai akurasi dari model baseline. Kseimpulannya adalah model klasifikasi dengan decision tree menjadi model yang paling baik untuk melakukan posttager bahasa Indonesia dalam pengujian yang dilakukan.