Implementasi POS-Tagging Menggunakan Pendekatan Baseline dan HMM

Laporan Tugas 2 Pemrosesan Bahasa Alami

Oleh:
Irfan Dwi Prakoso (1301160164)
Rezza Nafi Ismail (1301160425)

Fakultas Informatika
Universitas Telkom
Bandung
2019

1. Rincian Pendekatan yang Digunakan

1.1. Baseline

Baseline merupakan salah satu pendekatan dalam POS-tagging yang menentukan tag sebuah kata berdasarkan tag yang paling sering muncul untuk kata tersebut pada data yang sudah memiliki tag. Baseline memiliki tingkat akurasi yang tinggi yaitu lebih dari 90%, karena hanya melihat tag yang sering muncul.

1.2. *Hidden Markov Model* (HMM)

Hidden Markov Model atau HMM merupakan pendekatan POS-tagging yang dapat digambarkan dengan diagram weighted FSA atau *Finite State Automata*. Pendekatan HMM tidak hanya melihat seberapa sering kata muncul sebagai tag apa namun juga melihat dari transisi dari setiap tag-nya. Implementasi HMM POS-tagging dapat dilakukan dengan menggunakan algoritma Viterbi.

2. Analisis Pengujian

Data set menggunakan corpus yang diambil dari alamat web bahasa.cs.ui.ac.id/postag/corpus. Data train diambil dari 1000 kalimat pertama dan data test diambil dari 20 kalimat setelahnya. Berikut merupakan hasil pengujian POS-tagging menggunakan pendekatan Baseline dan HMM :

Perbandingan Akurasi

Baseline

In [17]:
```
1 print('Include not found: ', baseline_acc_loss, '%')
2 print('Found only: ', baseline_acc, '%')
```

```
Include not found:  92.20779220779221 %
Found only:  96.88109161793372 %
```

HMM

In [18]:
```
1 print('Include not found: ', hmm_acc_loss, '%')
2 print('Found only: ', hmm_acc, '%')
```

```
Include not found:  80.89053803339517 %
Found only:  84.99025341130604 %
```

Dari hasil pengujian diatas, dapat dilihat bahwa pendekatan baseline memiliki akurasi yang lebih tinggi dari pada HMM. Namun, perlu diketahui bahwa corpus yang diambil sangat kecil dan belum merepresentasikan banyak kalimat yang sebenarnya dapat ditemukan di kehidupan nyata. Sehingga, pengecualian-pengecualian atau anomali yang mungkin muncul tidak ada di dalam corpus. Oleh karena itu, pendekatan baseline memiliki keunggulan yang jauh lebih besar dibanding pendekatan HMM untuk kasus ini karena hanya melihat dari kemunculan tag yang paling sering, dan bukan berarti bahwa HMM lebih buruk dibanding baseline.