LAPORAN TUGAS PEKAN 11 NLP : MACHINE TRANSLATION

CLARISA HASYA YUTIKA | 1301174256 | IF 41 GAB01

1. Deskripsi Masalah

Implementasi pelatihan alignment dari teks paralel korpus yang didefinisikan sendiri, terdiri dari 20 pasang kalimat. Teks paralel korpus tersebut berisi teks dengan bahasa sumber adalah Bahasa Indonesia, dan bahasa target adalah Bahasa Sunda

2. Hasil & Analisis

Setelah mendefinisikan 20 kalimat untuk teks parallel korpus dan alignment. Akan dilakukan pelatihan menggunakan IBMModel1. Akan dilakukan 2 eksperimen, yaitu eksperimen pertama akan dilakukan iterasi sebanyak 5 iterasi, dan eksperimen kedua akan dilakukan iterasi sebanyak 100 iterasi. Kemudian akan dihitung skor probability alignment

- 5 iterasi

```
com_ibm1 = IBMModel1(corpus, 5)
```

```
print(round(com_ibm1.translation_table["saya"]["abdi"], 3) )
```

```
0.959
```

```
print(round(com_ibm1.translation_table["kecil"]["leutik"], 3) )
```

```
0.626
```

```
print(round(com_ibm1.translation_table["besar"]["ageung"], 3) )
```

```
0.461
```

```
print(round(com_ibm1.translation_table["ini"]["ieu"], 3) )
```

```
0.799
```

```
print(round(com_ibm1.translation_table["itu"]["éta"], 3) )
```

```
0.461
```

```
print(round(com_ibm1.translation_table["baju"]["acuk"], 3) )
```

```
0.841
```

```
print(round(com_ibm1.translation_table["sampah"]["runtah"], 3) )
```

```
0.415
```

```
print(round(com_ibm1.translation_table["membuang"]["miceun"], 3) )
```

```
0.415
```

- 100 iterasi

```
com_ibm1 = IBMModel1(corpus, 100)
```

```
print(round(com_ibm1.translation_table["saya"]["abdi"], 3) )
```

```
1.0
```

```
print(round(com_ibm1.translation_table["kecil"]["leutik"], 3) )
```

```
0.993
```

```
print(round(com_ibm1.translation_table["besar"]["ageung"], 3) )
```

```
0.5
```

```
print(round(com_ibm1.translation_table["ini"]["ieu"], 3) )
```

```
1.0
```

```
print(round(com_ibm1.translation_table["itu"]["éta"], 3) )
```

```
0.5
```

```
print(round(com_ibm1.translation_table["baju"]["acuk"], 3) )
```

```
1.0
```

```
print(round(com_ibm1.translation_table["sampah"]["runtah"], 3) )
```

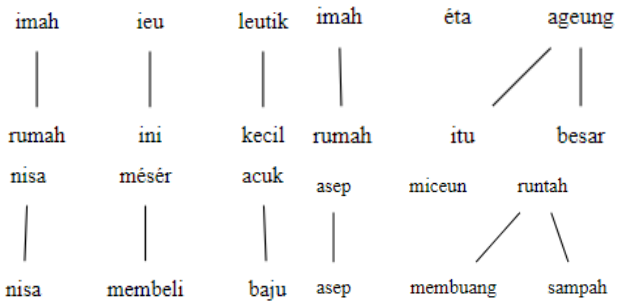
```
0.498
```

```
print(round(com_ibm1.translation_table["membuang"]["miceun"], 3) )
```

```
0.498
```

Dari hasil 2 eksperimen tersebut dilihat dari nilai skor probability alignment dengan proses pelatihan IBMModel1 sebanyak 100 iterasi, skor probabilitynya lebih bagus. Kata yang memiliki skor probabilitas tertinggi adalah ‘saya’, ‘abdi’, ‘ini’, ‘ieu’, ‘baju’, ‘acuk’, ‘kecil’, ‘leutik’. Sedangkan kata yang memiliki skor probabilitas rendah adalah ‘sampah’, ‘runtah’, ‘membuang’, ‘miceun’, ‘besar’, ‘ageung’,’itu’, ‘eta’. Kata yang mendapat skor probabilitas rendah dikarenakan terdapat kesalahan pada alignment.

Dari hasil pelatihan IBMModel1 masih terdapat alignment yang tidak sesuai. Hal ini dapat terjadi karena proses pencarian menyeluruh pada table translasi-frase untuk mendapatkan translasi frase dengan probability yang paling besar tidak praktis, sehingga banyak algoritma decoding mengaplikasikan beam search pruning (pada tiap iterasi, simpan most promising states, pangkas state lain).



Link Colab Tugas :
https://colab.research.google.com/drive/1hwv3L-jDmNr1Iec-bmQzRyXuGlwtDUsp?usp=sharing