Traitement du Langage Naturel et Linguistique

LE BELLEGO Victor, MARTZLOFF Alice, MOUSSOX Vincent

M2 IAAA 2021/2022

1 Certaines langues sont elles plus difficiles à analyser que d'autres ?

L'objectif de ce projet est d'analyser et de comprendre les raisons pour lesquelles des analyseurs syntaxiques partageant la même architecture et entraînés sur la même quantité de données obtiennent des performances très différentes sur différentes langues. On peut observer ce phénomène dans la Table 2 qui présente les performances calculées à l'aide des mesures LAS (Labeled Accuracy Score) et UAS (Unlabeled Accuracy Score) obtenues par un analyseur sur 36 langues différentes.

lang	las	uas
da	66.18	73.06
hr	58.88	69.67
id	69.99	75.63
ar	60.88	69.78
eu	47.94	58.17
it	74.82	81.01
fa	47.70	56.17
es	66.09	71.01
ro	57.23	68.61
de	58.12	64.19
hi	64.03	72.54
zh	46.76	55.60
lv	51.46	59.32
he	64.82	70.06
ko	46.06	55.00
ja	71.54	82.71
ca	64.79	71.13
en	66.57	71.22
pt	70.84	74.55
et	59.04	73.09
fr	68.40	73.46
el	69.96	76.58
pl	68.76	80.69
sv	64.52	71.11
cs	69.77	77.40
nl	54.92	63.69
hu	57.21	69.24
bg	68.08	78.14
vi	48.40	49.62
sl	49.46	61.51

../tbp-master/test.py

```
1  import os
2  from os import listdir
3  from subprocess import run
4  import pandas as pd
5  from tabulate import tabulate
```

```
6
7  os.chdir('/Users/victorlebellego/Documents/Dev/Python/IAAA/TLNL/tbp-master/')
8  list_lang = [k[6:8] for k in listdir('data') if '.conllu' in k and 'train' in k]
9  print('Training '+str(len(list_lang))+' languages')
10
11 os.chdir('expe/')
12 i=1
13 nbre = len(list_lang)
14 for lang in list_lang:
15     try :
16         print('Testing '+lang+' '+str(i)+'/'+str(nbre))
17         command = 'make lang=' + lang
18         run(command.split(), capture_output=True)
19         print('Success !')
20         i += 1
21     except :
22         print("Erreur " + lang)
23         pass
24
25 print('Reading results...')
26 os.chdir('out/')
27 df = pd.DataFrame(columns=["lang", "las", "uas"])
28 for lang in list_lang:
29     try:
30         with open(lang+'.res','r') as f:
31             lines = f.readlines()
32             to_append = lines[-1].split()
33             df_length = len(df)
34             df.loc[df_length] = to_append
35     except :
36         print("Erreur " + lang)
37         pass
38
39 df.set_index('lang', inplace=True)
40 df.to_csv('out.csv')
41 print(tabulate(df, headers='keys', tablefmt='psql'))
```