typo

Inspection of Ten UD Corpora Using UdPipe Parser

L-545

Practical 04

Ali Aljubailan

April 10, 2019

Ten UD treebanks were inspected to detect potential errors that might be found with UdPipe parser. Those treebanks were:

- UD Arabic PADT
- UD Basque BDT
- UD Coptic Scriptorium
- UD Danish DDT
- UD Finnish TDT
- UD Hebrew HTB
- UD Japanese GSD
- UD Latin ITTB
- UD Latvian LVTB
- UD Urdu UDTB
- UD Uyghur UDT

These copora were intentionally picked up from various language families, in order to test whetehr UdPipe parser's efficiency, as a language-agnostic parser, could be affected with certain features of languages. However, probably of the most obvious observation were the wide range in difference between languages in the average of UAS (Unlabelled attachment Score) and LAS (Labelled Attachment Score); i.e., in certain languages LAS and UAS were extremely low, in others they fairly were on average, and in other languages the two scores were obsevably higher. In details, Coptic was found to score the lowest LAS and UAS with merely 8.48 out of 100, while UAS was 25.32, which probably means that this Coptic UD corpus lacks efficient parsing treatment. Thus that might call for more work on this corpus for better parsing accuracy and thus better natural language processing results.

```
Metrics     | Precision |    Recall |  F1 Score | AligndAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |    100.00 |    100.00 |    100.00 |    100.00
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |    100.00 |    100.00 |    100.00 |    100.00
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |     25.32 |     25.32 |     25.32 |     25.32
LAS         |      8.48 |      8.48 |      8.48 |      8.48
Alis-MacBook-Air:udpipe alialjubailan$
```

The poor UAS and LAS score of Coptic and Uygur respectively may be more obvious when compared not only with hi-score languages such as Japanese, but it also can be obvious by comparing with the second lowest-UAS corpus, which is Uygur, as in the figure below:

```
u
Metrics     | Precision |    Recall |  F1 Score | AligndAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |    100.00 |    100.00 |    100.00 |    100.00
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |    100.00 |    100.00 |    100.00 |    100.00
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |     73.12 |     73.12 |     73.12 |     73.12
LAS         |     58.00 |     58.00 |     58.00 |     58.00
Alis-MacBook-Air:udpipe alialjubailan$ python3 ./evaluation_script/c
onllu
```

It can be seen that although Uygur is probably beyond the expected level, it still achieves a widely higher score than Coptic, almost by tripled times for UAS and double-tripled times for LAS.

The next fairly averaged score languages can be categorized into two main groups, based on the score they achieved. For instance, Urdu, Latin, and Danish respectively can be set as one group that achived an overall UAS Score of 85 out of 100 (where Urder is highest among this group with 87.85/100), and an overall LAS score of 81.87/100, where Latin is the highest LAS score in this group

with 82.88/100. Below are the tables of the three languages in this group respectively based on the highest scores, where Danish is the lowest in both UAS and LAS in this set.

```
Metrics    | Precision |    Recall |  F1 Score | AligndAcc
-----------+-----------+-----------+-----------+-----------
Tokens     |    100.00 |    100.00 |    100.00 |
Sentences  |    100.00 |    100.00 |    100.00 |
Words      |    100.00 |    100.00 |    100.00 |
UPOS       |    100.00 |    100.00 |    100.00 |    100.00
XPOS       |    100.00 |    100.00 |    100.00 |    100.00
Feats      |    100.00 |    100.00 |    100.00 |    100.00
AllTags    |    100.00 |    100.00 |    100.00 |    100.00
Lemmas     |    100.00 |    100.00 |    100.00 |    100.00
UAS        |     87.85 |     87.85 |     87.85 |     87.85
LAS        |     82.80 |     82.80 |     82.80 |     82.80
Alis-MacBook-Air:udpipe alialjubailan$ python3 ./evaluation_script/
```

Urdu Scores

```
Metrics    | Precision |    Recall |  F1 Score | AligndAcc
-----------+-----------+-----------+-----------+-----------
Tokens     |    100.00 |    100.00 |    100.00 |
Sentences  |    100.00 |    100.00 |    100.00 |
Words      |    100.00 |    100.00 |    100.00 |
UPOS       |    100.00 |    100.00 |    100.00 |    100.00
XPOS       |    100.00 |    100.00 |    100.00 |    100.00
Feats      |    100.00 |    100.00 |    100.00 |    100.00
AllTags    |    100.00 |    100.00 |    100.00 |    100.00
Lemmas     |    100.00 |    100.00 |    100.00 |    100.00
UAS        |     85.55 |     85.55 |     85.55 |     85.55
LAS        |     82.88 |     82.88 |     82.88 |     82.88
Alis-MacBook-Air:udpipe alialjubailan$
```

Latin Scores

```
Metrics     | Precision |    Recall |  F1 Score | AligndAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |    100.00 |    100.00 |    100.00 |    100.00
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |    100.00 |    100.00 |    100.00 |    100.00
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |     83.06 |     83.06 |     83.06 |     83.06
LAS         |     79.95 |     79.95 |     79.95 |     79.95
Alis-MacBook-Air:udpipe alialjubailan$
```

Danish Scores

The second group of this categorization entails Latvian, Arabic, and Basquese respectively based on the highest score, which is achieved by Latvian (79.81/100 for LAS and 82.63 for UAS). However, difference in score in this group languages seems to be the lowest among all languages. That is, those three languages differ from each other particularly in LAS by less than 3 points between the highest and lowest. In comparison with the previous three languages (Urdu, latin, and Danish), the difference in LAS between the highest in that group, Urdu, and the lowest, Danish, almost touched 5 points. Hence, the close LAS score in the Latvian, Arabic, and Basquese group might be interesting to investigate explanations as each langauge of those is categorized under a different language family, which could probably result in factors either about the parser itself or the three languages and other sinilar cases that might be able to enhance the work of parsing in general. Below are listed the three language scores respectively based on the highest:

```
Alis-MacBook-Air:udpipe alialjubailan$ python3 ./evaluation_script/c
onllu
Metrics     | Precision |    Recall |  F1 Score | AligndAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |    100.00 |    100.00 |    100.00 |    100.00
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |    100.00 |    100.00 |    100.00 |    100.00
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |     82.63 |     82.63 |     82.63 |     82.63
LAS         |     79.02 |     79.02 |     79.02 |     79.02
Alis-MacBook-Air:udpipe alialjubailan$
```

Latvian Score

```
Alis-MacBook-Air:udpipe alialjubailan$ python3 ./evaluation_scri
ADT/ar_padt-ud-test.conllu ar_PADT-testout.conllu
Metrics     | Precision |    Recall |  F1 Score | AligndAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |    100.00 |    100.00 |    100.00 |    100.00
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |    100.00 |    100.00 |    100.00 |    100.00
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |     79.81 |     79.81 |     79.81 |     79.81
LAS         |     76.01 |     76.01 |     76.01 |     76.01
Alis-MacBook-Air:udpipe alialjubailan$
```

Arabic PADT

```
Metrics     | Precision |    Recall |  F1 Score | AligndAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |    100.00 |    100.00 |    100.00 |    100.00
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |    100.00 |    100.00 |    100.00 |    100.00
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |     79.64 |     79.64 |     79.64 |     79.64
LAS         |     75.35 |     75.35 |     75.35 |     75.35
Alis-MacBook-Air:udpipe alialjubailan$
```

Basquese Score

Japanese enjoys the highest score among the whole inspected set, with a UAS of 94.15/00 and a LAS of 92.82. This parsing accuracy rate is fairly far probably from even closest language of all languages. It is also important to notice that in Japanese the difference between LAS and UAS is only 1.33 point, which is interesting compared to the other inspected languages, in which the difference between LAS and UAS arrange from approximately 13 points (in Coptic) to approximately 3 points for Latin, that has the closest difference rate after Japanese.

```
Alis-MacBook-Air:udpipe alialjubailan$ python3 ./evaluation_scr
-GSD/ja_gsd-ud-test.conllu ja_gsd-ud-testoutput.conllu
Metrics     | Precision |    Recall |  F1 Score | AligndAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |    100.00 |    100.00 |    100.00 |    100.00
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |    100.00 |    100.00 |    100.00 |    100.00
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |     94.15 |     94.15 |     94.15 |     94.15
LAS         |     92.82 |     92.82 |     92.82 |     92.82
Alis-MacBook-Air:udpipe alialjubailan$
```

Japanese_Score

In sum, observably all languages in this experiment achieved the same score for all parameters except UAS and LAS. They all scored 100 in parameters such as UPOS, XPOS, Lemmas, and Tokens. However, a wide range of different UAS and LAS rates going from very low to high can be observed. As languages were picked up from various language families, there stands a question whether Udpipe parser's efficiency is merely affected by a language features or there might be other certain factors that considerably play a role in this efficiency.