Dependency Parsing

Amritanshu Singh

Report

The Universal Dependency dataset used for this practical is the French ParTUT corpus with 27648 tokens, consisting of 1020 phrases and 28586 syntactic terms. There are 4152 tokens (15%) in this corpus that are not followed by a space. There are 1 category of words with spaces in this corpus. As mentioned in the practical, 10 trees were inspected with this corpus using the parser.

```
Iteration 1: training logprob -3.4103e+04
Iteration 2: training logprob -4.4898e+04
Iteration 3: training logprob -3.1865e+04
Iteration 4: training logprob -2.3122e+04
Iteration 5: training logprob -1.8354e+04
Iteration 6: training logprob -1.4683e+04
Iteration 7: training logprob -1.2715e+04
Iteration 8: training logprob -1.1741e+04
Iteration 9: training logprob -1.1711e+04
Iteration 10: training logprob -1.2048e+04
```

Following are the outcome of the evaluation metrics for the UDpipe parser for this corpus:

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	88.90	88.90	88.90	88.90
LAS	87.05	87.05	87.05	87.05

As observed from the metrics listed above, for this corpus, UDpipe parser performs well with UAS and LAS scoring 88.90 and 87.05 respectively.

Most of the sentences has been parsed correctly but in some cases part-of-the-speech tagger is incorrect.

For example, in the sentence “Paternité-Partage des conditions initiales à l'identique 2.0.” , ‘Partage’ is marked as noun instead of verb which is incorrect.

References

1. Zeldes, Amir. (2017). The GUM corpus: creating multilayer resources in the classroom. Language Resources and Evaluation. 51. 581-612. 10.1007/s10579-016-9343-x.
2. https://universaldependencies.org/treebanks/fr_partut/index.html