Report Tokenization

In this assignment I worked with 4 languages: Cantonese, Upper Sorbian, Buryat, and Udmurt.

Task 1.

The task is about checking the accuracy of existing models used for low-resource languages

./udpipe --tokenize model < input_data > tokenized_data

1) *Cantonese (UD test.conllu)*: for tokenization of Cantonese, I chose Chinese model, because it is the language from Chinese branch.

Model: chinese-gsd-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	13901	13918	77.04%	76.95%	77.00%
Sentences	992	1004	74.29%	73.41%	73.85%

Model: chinese-gsdsimp-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	15851	13918	64.17%	73.09%	68.34%
Sentences	962	1004	77.23%	74.00%	75.58%

2) *Upper Sorbian (UD test.conllu)* is a West Slavic language, in the same branch with Czech, Polish, and Slovak. Therefore, I decided to try all models of these 3 languages for tokenization.

- Czech models

Model: czech-pdt-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10701	10736	99.55%	99.23%	99.39%
Sentences	627	623	92.66%	93.26%	92.96%

Model: czech-cac-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10547	10736	98.55%	96.81%	97.67%

Sentences	886	623	54.63%	77.69%	64.15%

Model: czech-cltt-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10176	10736	96.06%	91.05%	93.49**%**
Sentences	908	623	42.07%	61.32%	49.90**%**

Model: czech-fictree-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10652	10736	99.44%	98.66%	99.05%
Sentences	737	623	70.69%	83.63%	76.62%

- Polish models

Model: polish-pdb-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10591	10736	99.26%	97.92%	98.59%
Sentences	643	623	89.27%	92.13%	90.68%

Model: polish-lfg-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10499	10736	98.47%	96.29%	97.37%
Sentences	948	623	46.62%	70.95%	56.27%

- Slovak model

Model: slovak-snk-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10475	10736	98.67%	96.27%	97.46%
Sentences	724	623	68.92%	80.10%	74.09%

All models show high accuracy for words, but some of them are rather bad for sentence. The best one is czech-pdt-ud-2.5-191206.udpipe

3) *Buryat (UD test.conllu)* is a language of Mongolic family. However, there is no model for Mongolian language, so I decided to try tokenizing it with Russian models, just because it is spoken in some Russian regions and uses Cyrillic writing system.

Model: russian-gsd-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	9955	10032	97.74%	96.99%	**97.36%**
Sentences	846	908	93.38%	87.00%	**90.08%**

Model: russian-syntagrus-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	9969	10032	97.29%	96.68%	96.99%
Sentences	849	908	93.52%	87.44%	90.38%

Model: russian-taiga-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10074	10032	98.72%	99.13%	98.93%
Sentences	969	908	79.67%	85.02%	82.26%

All models show rather accuracy for both words and sentences.

4) *Udmurt* language is a language of Uralic family, of Finno-Ugric group. Therefore, there are 3 ways to tokenize this language: with Finnish models, with Estonian models, or with Hungarian models. There is no data in UD for that language, so I've created a dataset by myself with Declaration of Human Rights. There is no gold data, so I was trying to evaluate the results just looking at it.

Model: finnish-ftb-ud-2.5-191206.udpipe.

Accuracy check:
Doesn't look good for sentences, ok for punctuation, ok for words.

Model: finnish-tdt-ud-2.5-191206.udpipe.

Accuracy check:
Doesn’t look good for sentences, has problems with punctuation, ok for words.

Model: estonian-edt-ud-2.5-191206.udpipe.

Accuracy check:
Doesn’t look good for sentences, ok for punctuation, ok for words.

Model: estonian-ewt-ud-2.5-191206.udpipe.

Accuracy check:
Looks a bit better for sentences, ok for punctuation, ok for words.

Model: hungarian-szeged-ud-2.5-191206.udpipe.

Accuracy check:
Looks a bit better for sentences, ok for punctuation, ok for words.

It’s hard to say which model is the best one.

Task 2.

The task is about training the tokenizers for low-resource languages and compare their accuracy with accuracy of existing models.

./udpipe --train --tagger=none --parser=none model.udpipe < train_data.conllu

./udpipe --tokenize --accuracy model.udpipe < test_data.conllu

1) Cantonese

I trained the tokenizer using Cantonese data from UD dataset, splitting test.conllu file to train and test parts

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	4569	4417	83.94%	86.82%	85.35%
Sentences	206	205	84.95%	85.37%	85.16%

Resulting accuracy is better than in Chinese models.

2) Upper Sorbian

I trained the tokenizer using Upper Sorbian data from UD dataset (train.conllu, test.conllu)
Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10694	10736	98.68%	98.30%	98.49%
Sentences	723	623	65.70%	76.24%	70.58%

Resulting accuracy is worse than for the best Czech model.

3) Buryat

I trained the tokenizer using Buryat data from UD dataset (train.conllu, test.conllu)

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	10014	10032	96.69%	96.52%	96.61%
Sentences	896	908	92.52%	91.30%	91.91%

Resulting accuracy is comparable with Russian models.