Where␣are␣we␣Still␣Split␣on␣Tokenization␣?

Task definition

Input:
If␣momma␣ain't␣happy,␣nobody␣ain't␣happy.

Tokenization:
If␣momma␣ain't␣happy␣,␣nobody␣ain't␣happy␣.

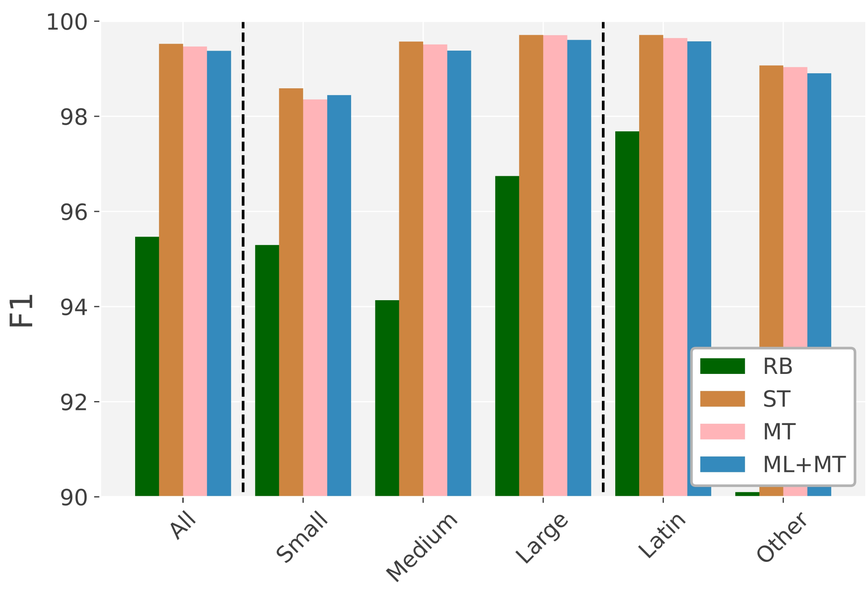
Multi-word expansions:
If␣momma␣is␣not␣happy,␣nobody␣is␣not␣happy.

Subword segmentation:
If␣mo␣##mma␣ai␣##n␣'␣t␣happy␣,␣no␣##body␣ai␣##n␣'␣t␣happy␣.

"specialization" should be tokenized!

You $%@!!, you mean subword segmented!

Results



Ah, rulebased definitely doesn't cut it anymore, and train-datasize and script are important!

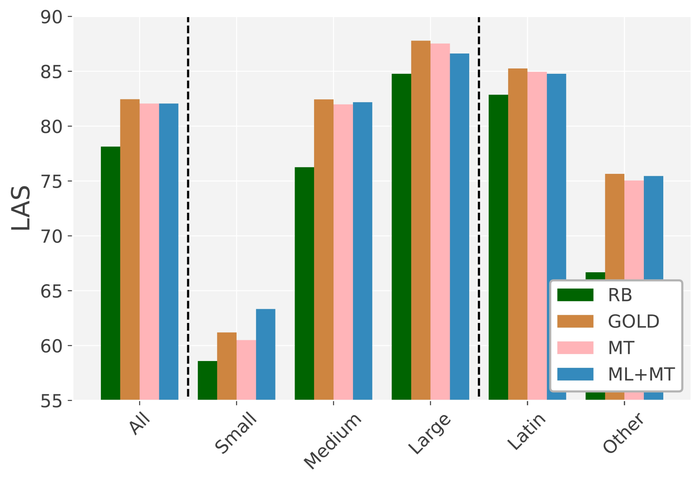
And look, multi-task and multi-lingual learning are not detremental!

Test-only treebanks:

setting	F1 tok.	F1 LAS	# treebanks
all	93.23	38.72	90
in-language	95.11	68.20	34
in-script	94.16	40.45	84
new-script	80.11	14.41	6

Performance is in general much lower for cross-dataset setups. For new scripts the drop is even ~15 points!

Downstream results (dep. parsing)



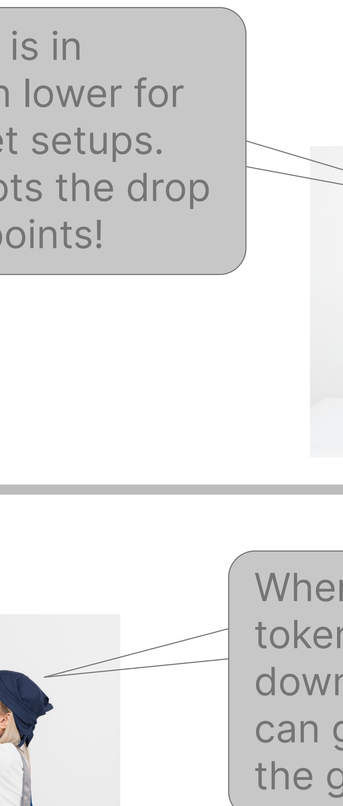
When using SOTA tokenization for a downstream task, we can get quite close to the gold tokenization

Qualitative analysis

* Unknown subwords:
  - Script
  - Emojis
* Adpositions
* Challenging cases: is there anyway

* Compound words
* Names that consists of lexical tokens

MC␣Donalds
f/2.77
2-0



MC␣Donald␣s
f␣/␣2.7
2␣-␣0

IT UNIVERSITY OF COPENHAGEN

Implemented in MaChAmp:
https://github.com/machamp-nlp/machamp