Distilling the Knowledge of Romanian BERTs Using Multiple Teachers

Andrei-Marius Avram[1], **Darius Catrina**[1], **Dumitru-Clementin Cercel**[2], **Mihai Dascalu**[2], **Traian Rebedea**[2], **Vasile Păiș**[1], **Dan Tufiș**[1]

[1]Research Institute for Artificial Intelligence, Romanian Academy,
[2]University Politehnica of Bucharest, Faculty of Automatic Control and Computers

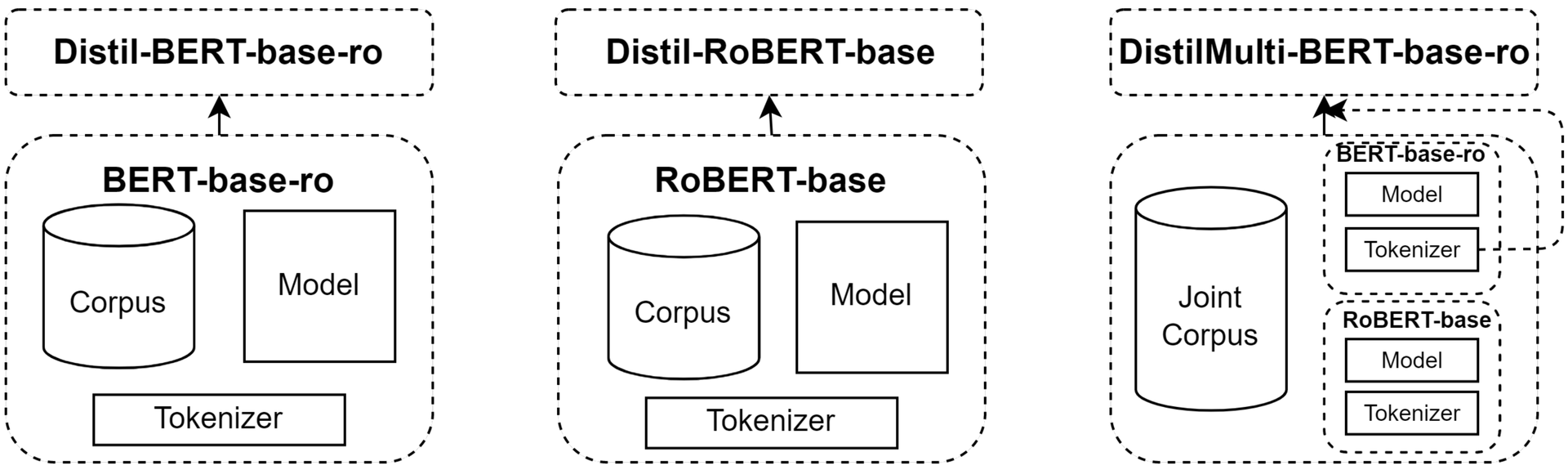
Motivation

- Running large-scale pre-trained language models in computationally constrained environments is a challenging problem
- Model optimizations (e.g., distillation, quantization) focus mostly on the English language, thus widening the gap when considering low-resource languages
- As a response, we introduce and evaluate three distilled models for the Romanian language:
  1. Distil-BERT-base-ro
  2. Distil-RoBERT-base
  3. DistilMulti-BERT-base-ro

Task Evaluation

- *Part-of-Speech Tagging*: Universal Part-of-Speech (UPOS) and eXtended Part-of-Speech (XPOS)
- Named Entity Recognition (NER): Tag a sequence of tokens with IOB entities
- *Sentiment Analysis*: Positive or negative sentiment (SAPN) or rating (SAR)
- *Dialect Identification* (DI): Identify the Romanian/Moldavian dialects in news articles
- *Semantic Textual Similarity* (STS): Given a pair of sentences, measure their semantically relatedness

Distillation Process



Results

Model	UPOS	XPOS	NER	SAPN	SAR	DI	STS
BERT-base-ro	98.00	96.46	**85.88**	97.94	79.61	95.41	79.94
RoBERT-base	98.02	**97.15**	85.14	**98.20**	79.40	96.17	81.18
Distil-BERT-base-ro	97.97	97.08	83.35	98.12	**80.51**	**96.31**	80.57
Distil-RoBERT-base	97.12	95.79	83.11	97.61	79.58	96.11	79.80
DistilMulti-BERT-base-ro	**98.07**	96.93	83.22	97.74	79.77	96.18	**80.66**

Loyalty

Regression loyalty (R-L): the Pearson Coefficient between the outputs of the teacher and the student

Model	L-L	P-L	R-L
BERT-base-ro	87.80	**74.76**	**94.96**
RoBERT-base	84.63	71.95	92.24
Distil-BERT-base-ro	**89.75**	73.23	94.64

Conclusions

- Introduced three distilled models for the Romanian language that are approximately 35% smaller and that maintain most of the performance of their teachers
- Evaluated these distilled models on five Romanian task in comparison to their original teachers
- Introduced a new metric that traces how well the student and teacher predictions are aligned in the regression tasks – *regression loyalty*

Acknowledgments: This research was supported by the EC grant INEA/CEF/ICT/A2018/28592472 for the Action No: 2019-EU-IA-0034 entitled “Curated Multilingual Language Resources for CEF.AT” (CURLICAT) and by the Romanian Ministry of European Investments and Projects through the Competitiveness Operational Program (POC) project "HOLOTRAIN" (grant no. 29/221, SMIS code: 129077).