The MaCoCu project: Massive Collection, Curation and Evaluation of Monolingual and Bilingual Data

Rik van Noord and Antonio Toral

rijksuniversiteit groningen

Universitat d'Alacant
Universidad de Alicante

Jožef Stefan Institute

MaCoCu

Create corpora for under-resourced EU languages

- Bulgarian, Croatian, Icelandic, Maltese, Slovene and Turkish

How: Crawling respective top-level domains

- No need to rely on Common Crawl
- Focus on cleaning and deduplicating
- Main advantage: more data!

	Parallel with English		Monolingual	
Language	**# Sents**	**# Words**	**# Sents**	**# Words**
Bulgarian	3.9M	159M	226M	3509M
Croatian	3.1M	135M	140M	2318M
Icelandic	0.4M	14.4M	42M	645M
Maltese	1.2M	70M	26M	348M
Slovene	3.2M	137M	117M	1779M
Turkish	10.3M	513M	130M	4346M

Goal ➡ Evaluating the quality of these corpora

How ➡ Training NMT systems and large LMs

Parallel Data Evaluation

Experiment

- Train strong NMT models with and without MaCoCu

COMET on Flores (devtest)	bg	hr	is	mt	sl	tr
Data set without MaCoCu	13.9M	5.1M	3.2M	2.2M	10.9M	4.9M
Best model without MaCoCu	79.2	80.6	47.0	80.7	77.0	81.3
Best model with MaCoCu	**+0.9**	**+0.5**	**+2.4**	**+1.0**	**+0.3**	**-1.2**

What is going on with Turkish?

Turkish (COMET)	FL	W16	W17	W18	TED	Wiki	QED
Without MaCoCu	81.3	72.6	77.3	74.0	62.9	61.1	51.1
With MaCoCu	**-1.2**	**+2.3**	**+0.4**	**+2.2**	**+1.6**	**+1.0**	**+0.8**

Human Evaluation with professional translators

	bg	hr	is	mt	sl	tr
Without MaCocu preferred (%)	9.2	25.7	22.9	26.1	37.4	36.2
Same quality (%)	79.8	43.9	49.5	36.2	22.9	22.3
With MaCoCu preferred (%)	11.0	30.5	27.6	37.7	39.6	41.5
Relative with/without difference	**+1.8**	**+4.8**	**+4.7**	**+11.6**	**+2.2**	**+5.3**

Monolingual Data Evaluation

Goal ➡ training large language models

BERTovski (bg/mk) and **MaltBERT** (mt)

- ➡ Train RoBERTa model from scratch
- ➡ bg/mk ➡ 300k steps, 74GB text, 3 TPU weeks
- ➡ mt ➡ 100k steps, 3.2GB text, 4 TPU days

Why start from scratch?

- ➡ Continue training from XLM-R-large
- ➡ Training is slower, but converges a lot faster
- ➡ Icelandic & Turkish only trained on MaCoCu

All models are available through HuggingFace

	Bulgarian			Macedonian		
	UPOS	**XPOS**	**NER**	**UPOS**	**XPOS**	**NER**
XLM-R (baseline)	99.4	98.1	93.3	98.5	97.4	94.8
BERTovski (100k)	**-0.8**	**-0.5**	**-0.3**	**-0.6**	**-1.2**	**-1.7**
BERTovski (200k)	**-0.4**	**-0.3**	**0.0**	**-0.3**	**-1.2**	**-0.3**
BERTovski (300k)	**-0.3**	**-0.3**	**0.0**	**-0.4**	**-1.2**	**+0.5**
XLM-R-cont (67.5k)	**+0.1**	**+0.7**	**+1.8**	**+0.2**	**+0.3**	**+0.9**

Maltese	UPOS	XPOS
XLM-R (baseline)	94.5	95.0
MaltBERT (100k)	**+1.3**	**+0.8**
XLM-R-cont (50k)	**+3.5**	**+3.2**

	Icelandic			Turkish		
	UPOS	**XPOS**	**NER**	**UPOS**	**XPOS**	**NER**
XLM-R (baseline)	96.7	94.7	89.2	89.4	90.6	93.5
XLM-R-cont (75k/70k)	**+0.6**	**+0.6**	**+4.2**	**+0.1**	**+0.3**	**+0.7**

What's next?

Language Models

- ➡ Continue training BERTovski
- ➡ Continue from monolingual LMs instead of XLM-R
- ➡ Training smaller XLM-R-based models

Next release: June 2023

- ➡ Serbian, Montenegrin, Macedonian, Albanian, Ukrainian, Catalan

This action has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

Co-financed by the Connecting Europe Facility of the European Union

rikvannoord@gmail.com rikvannoord.nl & macocu.eu github.com/macocu/LanguageModels @RikVanNoord