Creating, Evaluating and Enriching Parallel Corpora in the MaCoCu Project

Rik van Noord and Antonio Toral

MaCoCu

Massive Collection and Curation of Mono-lingual and Bi-lingual Data

Create corpora for under-resourced EU languages

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

How: Crawling top-level domains per language

➡ No need to rely on Common Crawl

➡ Focus on cleaning and deduplicating

➡ Main advantage: more data!

Statistics of first release

	Parallel with English		Monolingual	
Language	**Sents**	**Words**	**Sents**	**Words**
Bulgarian	3.9M	159M	226M	3509M
Croatian	3.1M	135M	140M	2318M
Icelandic	0.35M	14.4M	42M	645M
Maltese	1.2M	70M	26M	348M
Slovene	3.2M	137M	117M	1779M
Turkish	10.3M	513M	130M	4346M

Our goals

➡ Evaluate the quality of the parallel corpora

➡ Enrich parallel corpora by determining translation direction

Evaluation by Machine Translation

Assumption

➡ Higher quality data leads to better MT systems

Experiment

➡ Train strong NMT models with and without MaCoCu

Data

Model Transformer-base (Marian)

Training CC-Aligned, ParaCrawl, Tilde and MaCoCu

Evaluation Flores, TED, WikiMatrix, QED

Metrics BLEU, COMET, etc

RQ1: does adding MaCoCu data improve a strong baseline?

COMET on Flores (devtest)	bg	hr	is	mt	sl	tr
Data set size 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**

Similar results across evaluation sets and metrics

➡ MaCoCu data clearly helps, except for Turkish

RQ2: Is MaCoCu of higher quality than ParaCrawl?

Experiment

Train models with same data set sizes, either based on sentences or bytes

COMET on Flores (devtest)	bg	hr	mt	sl
MaCoCu only	68.4	72.9	80.4	67.2
ParaCrawl (sentences)	**+2.9**	**+1.4**	**-0.3**	**-2.9**
ParaCrawl (bytes)	**+5.8**	**+0.8**	**-0.4**	**-1.2**

Too little data for Icelandic, Turkish is not in ParaCrawl

Enriching the Corpora

Goal: for each sentence pair, determine the original text

Motivation: support corpus translation research

How: fine-tuning a multi-lingual LM (XLM-R)

This is a very **challenging** task. Try it yourself:

Feyenoord heeft donderdag een fenomenale prestatie geleverd door ten koste van Olympique Marseille de finale van de Conference League te bereiken.

Feyenoord delivered a phenomenal performance on Thursday by reaching the final of the Conference League at the expense of Olympique Marseille.

Problem & Opportunity

➡ Not many parallel corpora have this information available

Training

➡ Our own annotated data for Croatian and Slovene

➡ Europarl for 20 langs, including Slovene and Bulgarian

➡ Downsample original English data

➡ Multi-lingual LM: apply **zero-shot classification**

Lang	Train data	Eval data	Accuracy
bg	Europarl (20) + MaCoCu	Europarl (bg)	81.5%
hr	MaCoCu	MaCoCu	87.2%
is	Europarl (20) + MaCoCu	WMT	70.0%
mt	Europarl (20) + MaCoCu	News	54.7%
sl	MaCoCu	MaCoCu	80.0%
tr	Europarl (20) + MaCoCu	WMT	79.1%

Domain bias

➡ Can be problematic for general domains

➡ But: a text about Dutch soccer is likely to be original Dutch

➡ In practice, it's better to take domain into account

What's next?

MaCoCu will be at LREC, EAMT, BUCC (LREC) and HumEval (ACL)

Soon: new pre-trained LMs for current languages

Next release: June 2023

➡ More data + Serbian, Montenegrin, Macedonian, Albanian

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/RikVN/ @rikvannoord