UniDive
2nd General Meeting
University of Naples "L'Orientale
Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

An Empirical Study of Multilingual Representations from Language Modeling and Translation

Shaoxiong Ji [1] **Timothee Mickus** [1] **Vincent Segonne** [2]
Alessandro Raganato [3] **Jörg Tiedemann** [1]
[1] University of Helsinki [2] Universite Grenoble Alpes [3] University of Milano-Bicocca
firstname.lastname@helsinki.fi

Work in progress

- a principled standpoint and train comparable MT and LM systems to contrast their cross-lingual and monolingual downstream performances;
- an empirical study on publicly available pretrained LM and MT systems and study whether continued training on MT helps or hinders the emergence of cross-lingual capabilities.

- Data
  - UNPC (Ziemski et al., 2016)
  - OpenSubtitles (Tiedemann, 2012)
- Languages: Arabic, Chinese, English, French, Russian, and Spanish
- Models
  1. Masked Language Modeling (MLM) with the BERT architecture (Devlin et al., 2019);
  2. Causal Language Modeling (CLM) with the GPT-2 architecture (Radford et al., 2019);
  3. Translation Language Modeling (TLM) with the GPT-2 architecture, where the input is the concatenation of a language pair following a setup similar to Conneau and Lample (2019);
  4. Denoising Sequenece-to-Sequence Langauge Modeling with BART architecture (Lewis et al., 2020);
  5. Machine Translation (MT) with the classic encoder-decoder transformer architecture (Vaswani et al., 2017) and the BART architecture (Lewis et al., 2020).

UNIVERSITY OF HELSINKI

Preliminary Results

	Model	Tasks							
		NC	XNLI	PAWS-X	QAM	QADSM	WPR	NER	POS
	mBERT	81.3	65.2	86.6	64.6	63.1	74.4	77.5	76.0
LM	XLM-R	**82.1**	**73.5**	88.9	67.4	**66.9**	**75.3**	**78.7**	**79.7**
	mBART	82.1	67.6	**89.2**	**67.8**	65.5	74.7	77.7	72.7
MT	NLLB 600M	76.0	68.3	73.4	61.5	63.9	73.7	54.2	71.4
	mBART m2o	80.4	65.9	85.6	63.9	63.9	73.7	61.5	70.8
CP	mBART o2m	65.4	48.1	81.7	58.4	62.7	73.2	55.1	55.7
	mBART m2m	78.3	60.2	87.2	63.2	62.8	73.7	71.9	69.7

Table: Average performance on cross-lingual tasks. We use the base architecture for mBERT and XLM-R. mBART scores are derived from the 12-layer encoder.