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

Abstract

Large language models (LMs) display impressive performances and have captured the attention of the NLP community. In this article, we focus on their cross-lingual generalization capabilities: We argue that machine translation (MT) systems ought to provide a reasonable comparison point, as they are expected to produce language-agnostic representations. We summarize two sets of experiments: First, we adopt a principled standpoint and train comparable MT and LM systems to contrast their cross-lingual and monolingual downstream performances; Second, we focus on publicly available pretrained LM and MT systems and study whether continued training on MT helps or hinders the emergence of cross-lingual capabilities.

Relevant UniDive working groups: WG3

1 Introduction

The ability of pretrained language to generalize across languages has been an active area of studies, with works ranging from linking typological factors to cross-lingual performances (Lin et al., 2019; Chai et al., 2022), to highlighting its benefits for specific tasks such as text processing, sentiment analysis or summarization (Xu et al., 2022; Wang et al., 2022). The successes of multilingual pretrained language models (LM) on cross-lingual tasks have been underscored time and time again (Wu and Dredze, 2019, e.g.,), and appears all the more surprising that they are often simply pretrained on datasets comprising multiple languages, without explicit cross-lingual supervision (cf. for instance Liu et al., 2020; although explicit supervision also exists, Xue et al., 2021). Explicit alignments such as linear mapping (Wang et al., 2019) and L2 alignment (Cao et al., 2020) between source and target languages do not necessarily improve the quality of cross-lingual representations (Wu and Dredze, 2020).

This is somewhat at odds with expectations from earlier studies in machine translation (MT). In particular, MT systems have had a historical connection with the concept of an interlingua—a language-independent representation space that MT systems can leverage to perform translation (Masterman, 1961; Lu et al., 2018). As such, MT models are expected to pick up on language-independent semantic features (Tiedemann, 2018)—though in practice, this shared representation space can be in a trade-off relationship with performance, which benefits from a greater separability of source language representations (Chen et al., 2023, e.g.). It should also be noted that previous studies have leveraged pretrained encoder-decoder LMs to built effective MT models (Liu et al., 2020; Tang et al., 2020): which suggests that MT and LM are not entirely unrelated tasks—although the evidence is conflicting here again (Vázquez et al., 2021).

Research questions In short, this state of affairs begs the question of whether MT systems do in fact learn some form of implicit cross-lingual alignment. This prompts us to study specifically how MT compares with multilingual LM when it comes to learning cross-lingual representations. More narrowly, we focus on verifying whether MT training objectives do favor the emergence of cross-lingual alignments more than LM objectives. We consider two separate but related approaches to answering this question: one where we adopt a principled perspective and learn strictly comparable models and contrast their cross-lingual performances, and one where we factor in the current state of the NLP research landscape, and study how existing publicly available MT models compare to publicly available LM systems on cross-lingual tasks.

Findings Our preliminary findings based on publicly available LM and MT models suggest that MT is not a good continued objective for pretrained multilingual LMs, as far as cross-lingual learning is concerned. However, those public models are trained with different corpora, and potential data contamination is a concern. We will conduct a more systematic analysis of the models trained with a more controlled setting, including training corpora, model architectures, and learning objectives.

2 Methods and settings

From a purely engineering-focused standpoint, the question of which of MT or LM is the most appropriate pre-training regimens for cross-lingual downstream application *a priori* is distinct from knowing which model one ought to work with in order to obtain higher performances for specific tasks. In practice, more resources might have been allocated to developing LM systems (or MT systems), making them a more appropriate starting point for cross-lingual tasks.

We start our inquiry by adopting a principled stance: We train strictly comparable models with MT and LM objectives before contrasting their performances on cross-lingual and mono-lingual tasks. We choose UNPC (Ziemski et al., 2016) and OpenSubtitles (Tiedemann, 2012) as the training corpora and consider six languages: Arabic, Chinese, English, French, Russian, and Spanish. To allow a systematic evaluation, we train models with various neural network architectures and learning objectives: (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). We have completed the training for MLM, CLM, TLM, and MT with a 6-layer encoder and 6-layer decoder. Other models are being trained.

3 Evaluation and preliminary results

We aim to evaluate models both publicly available and trained by us on various cross-lingual and monolingual NLP tasks. We start with cross-lingual tasks and plan to expand our evaluation to monolingual tasks. We will also evaluate machine translation performance and study the representation learned by different architectures and learning objectives once the model training has been completed. Here, we report some of our preliminary results.

3.1 Cross-lingual tasks and results

We consider cross-lingual NLP tasks, where model training for downstream applications is done in one language (usually English), and the trained model is evaluated in languages other than the language used for training. We use the XGLUE benchmark (Liang et al., 2020), a cross-lingual evaluation benchmark, and conduct our evaluation on natural language understanding tasks. The specific tasks consist of Named Entity Resolution (NER) (Sang, 2002; Sang and Meulder, 2003), Part of Speech Tagging (POS) (Zeman et al., 2020), News Classification (NC), XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019), Query-Ad Matching (QADSM), Web Page Ranking (WPR), and QA Matching (QAM). Table 1 shows the overall performance by averaging the scores of each language. XLM-R displays the highest performances on 6 out of 8 tasks, and mBART obtains the best average score on the last two. In most cases, models continually pretrained on MT (i.e., mBART m2o, mBART o2m, and mBART m2m) perform worse than language models (i.e., mBART).

	Model	Tasks							
		NC	XNLI	PAWS-X	QAM	QADSM	WPR	NER	POS
LM	mBERT	81.3	65.2	86.6	64.6	63.1	74.4	77.5	76.0
	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
CP	mBART m2o	80.4	65.9	85.6	63.9	63.9	73.7	61.5	70.8
	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 1: 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.

4 Conclusion

This proposal introduces our ongoing work. Our preliminary study on publically available models shows that continued training with machine translation models is beneficial for cross-lingual transfer. However, the preliminary study is based on models trained with different corpora. We will study a more controlled setting to fairly compare the performance of language and machine translation models and investigate the distributed representations learned by different models and learning objectives.

References

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4702–4712, Dublin, Ireland. Association for Computational Linguistics.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023. On the off-target problem of zero-shot multilingual neural machine translation. In *Findings of ACL*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Margaret Masterman. 1961. Semantic message detection for machine translation, using an interlingua. In *Proceedings of the International Conference on Machine Translation and Applied Language Analysis*, National Physical Laboratory, Teddington, UK.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. *ArXiv*, cs.CL/0209010.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *ArXiv*, cs.CL/0306050.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of LREC*, volume 2012, pages 2214–2218.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, volume 2084 of *CEUR Workshop Proceedings*, pages 188–197, Unknown. CEUR Workshop Proceedings. Digital humanities in the Nordic Countries DHN2018, DHN2018 ; Conference date: 07-03-2018 Through 09-03-2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Raúl Vázquez, Hande Celikkanat, Mathias Creutz, and Jörg Tiedemann. 2021. On the differences between BERT and MT encoder spaces and how to address them in translation tasks. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 337–347, Online. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A Survey on Cross-Lingual Summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual bert transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482.

Yuemei Xu, Han Cao, Wanze Du, and Wenqing Wang. 2022. A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations. *Data Science and Engineering*, 7(3):279–299.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *ArXiv*, abs/1908.11828.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, R Ziane, and et al. 2020. Universal dependencies 2.5.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).