A Comparison of Language Modeling and Translation as Multilingual Pretraining Objectives

Zihao Li, [1] Shaoxiong Ji, [1] Timothee Mickus, [1] Vincent Segonne, [2] and Jörg Tiedemann [1]
[1] University of Helsinki
[2] Université Bretagne Sud
firstname.lastname@{helsinki.fi, [2] univ-ubs.fr}

Abstract

Pretrained language models (PLMs) display impressive performances and have captured the attention of the NLP community. Establishing best practices in pretraining has, therefore, become a major focus of NLP research, especially since insights gained from monolingual English models may not necessarily apply to more complex multilingual models. One significant caveat of the current state of the art is that different works are rarely comparable: they often discuss different parameter counts, training data, and evaluation methodology.

This paper proposes a comparison of multilingual pretraining objectives in a controlled methodological environment. We ensure that training data and model architectures are comparable, and discuss the downstream performances across 6 languages that we observe in probing and fine-tuning scenarios. We make two key observations: (1) the architecture dictates which pretraining objective is optimal; (2) multilingual translation is a very effective pretraining objective under the right conditions. We make our code, data, and model weights available at https://github.com/Helsinki-NLP/lm-vs-mt.

1 Introduction

The release of BERT (Devlin et al., 2019) has marked a paradigm shift in the NLP landscape and has ushered in a thorough investment of the NLP research community in developing large language models that can readily be adapted to novel situations. The design, training, and evaluation of these models has become a significant enterprise of its own.

In recent years, that sustained interest has shifted also to encompass multilingual models (e.g., Muennighoff et al., 2022; Alves et al., 2024). There is considerable variation as to how such models are trained: For instance, some rely on datasets comprising multiple languages without explicit crosslingual supervision (e.g., Liu et al., 2020), and some use explicit supervision (Xue et al., 2021). One complication that arises from this blossoming field of study is that much of the work being carried out is not directly comparable beyond the raw performances on some well-established benchmark, a procedure which may well be flawed (Gorman and Bedrick, 2019). Avoiding apples-to-oranges comparison requires a methodical approach in strictly comparable circumstances, which is the stance we adopt in this paper.

In short, we focus on two variables- model architecture and pretraining objectives- and set out to train five models in strictly comparable conditions and compare their monolingual performances in three downstream applications: sentiment analysis, named entity recognition, and POS-tagging. The scope of our study spans from encoder-decoder machine translation models, to decoder-only causal language models and encoder-only BERT-like masked language models. We categorize them into double-stacks (encoder-decoder) and single-stacks (encoder-only or decoder-only) models. We intend to answer two research questions:

(i) Does the explicit cross-lingual training signal of translation objectives foster better downstream performances in monolingual tasks? (ii) Is the optimal choice of architecture independent of the training objective?

There are a prima facie reasons to favor either answers to both of these questions. For instance, the success of multilingual pretrained language models (LM) on cross-lingual tasks has been underscored repeatedly (Wu and Dredze, 2019, e.g.,), yet 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).

Our experiments provide tentative evidence that insofar as a BART denoising autoencoder architecture is concerned, models pretrained with a translation objective consistently outperform those trained with a denoising objective. However, for single-stack transformers, we observe causal language models to perform well in probing and masked language models to generally outperform translation and causal objectives when fine-tuned on downstream tasks. This leads us to conjecture that the optimal pretraining objective depends on the architecture. Furthermore, the best downstream results we observe appear to stem from a machine-translation system, highlighting that MT encoder-decoder systems might constitute an understudied but potentially very impactful type of pretrained model.

2 Methods and Settings

We start our inquiry by adopting a principled stance: We train strictly comparable models with MT and LM objectives before contrasting their performances on monolingual tasks.

2.1 Models and objectives

To allow a systematic evaluation, we train models with various neural network architectures and learning objectives. All models are based on the transformer architecture (Vaswani et al., 2017) and implemented in fairseq (Ott et al., 2019). We consider both double-stacks (encoder-decoder) and single-stacks (encoder-only or decoder-only) models.

The two double-stack models are variants of the BART architecture of (Lewis et al., 2020); they are trained either on a straightforward machine translation (MT) objective, using language tokens to distinguish the source, or on the original denoising auto-encoder objective of Lewis et al.. We refer to these two models as 2-LM and 2-MT respectively.

text[[127, 778, 472, 925], [496, 74, 839, 124]] We also consider three single-stack models: (i) an encoder-only model trained on the masked language modeling objective (MLM) of Devlin et al. (2019); (ii) an autoregressive causal language model (CLM), similar to Radford et al. (2019); and (iii) an autoregressive model trained to generate a sentence, followed by its translation in the language specified by a given control token, known as a translation language model (TLM) as proposed by Conneau and Lample (2019). We provide an example datapoint for each pretraining objective in Table 3, Appendix A.

2.2 Pretraining conditions

Our core focus is on guaranteeing comparable conditions across the different pretraining objectives we consider. This entails that our datasets need to be doubly structured: both in documents for CLM pretraining; and as aligned bitexts for MT pretraining. Two datasets broadly match these criteria: the UNPC (Ziemski et al., 2016) and OpenSubtitles (OpSub; Tiedemann, 2012) corpora. The choice also narrows down the languages considered in this study: we take the set of languages present in both resources, namely the six languages in UNPC: Arabic (AR), Chinese (ZH), English (EN), French (FR), Russian (RU), and Spanish (ES).

To guarantee that models are trained on the same data, whenever a document is available in multiple languages, we greedily assign it to the least represented language pair thus far and discard all other possible language pairs where it could have contributed; we then discard documents which cannot be used as bitexts. This ensures that all documents are used exactly

once for both document-level and bitext-level pretraining objectives. Dataset statistics are shown in Table 4, Appendix B.

To ensure a fair comparison, we control key variables, including tokenization (100k BPE pieces; Sennrich et al., 2016), number of transformer layers (12), hidden dimensions (512), attention heads (8), and feedforward layer dimensions (2048). We perform 600k steps of updates, using the largest batch size that fits into the GPU memory, deploy distributed training to make a global batch size of 4096, and apply the Adam optimizer (Kingma and Ba, 2017). Owing to the computational requirements, we only train one seed for each of the five types of models considered.

2.3 Downstream evaluation

The evaluations encompassed both sequence-level and token-level classification tasks using datasets tailored for sentiment analysis (SA), named entity recognition (NER), part-of-speech (POS) tagging, and natural language inference (NLI).

For SA, we utilized the Amazon review dataset (Hou et al., 2024) in English, Spanish, French, and Chinese. RuReviews (Smetanin and Komarov, 2019) for Russian, and ar_res_reviews (ElSahar and El-Beltagy, 2015) for Arabic. While the datasets for most languages were pre-split, ar_res_reviews required manual division into training, validation, and testing sets, using an 8:1:1 ratio.

For NER, we model the problem as an entity span extraction using a BIO scheme. In practice, we classify tokens into three basic categories: Beginning of an entity (B), Inside an entity (I), or Outside any entity (O). We use the MultiCoNER v2 dataset (Fetahu et al., 2023) for English, Spanish, French, and Chinese, MultiCoNER v1 (Malmasi et al., 2022) for Russian and the AQMAR Wikipedia NER corpus (Mohit et al., 2012a) for Arabic. Simplifying the NER task to these fundamental categories allows us to focus more on assessing the basic entity recognition capabilities of the models without the additional complexity of differentiating numerous entity types, which can vary significantly between languages and datasets.

For POS tagging, we utilized the Universal Dependencies (UD) 2.0 datasets (Nivre et al., 2020), selecting specific corpora tailored to each language to ensure both linguistic diversity and relevance. We select multiple UD treebanks per language, such that each language dataset comprises approximately 160,000 tokens, which are then split into training, validation, and testing segments with an 8:1:1 ratio.

For NLI, we employed the XNLI dataset (Conneau et al., 2018) for the six languages. The XNLI dataset consists of sentence pairs translated from the MultiNLI dataset (Williams et al., 2018) into 15 languages, providing consistent annotations across languages. The task focuses on classifying the relationship between pairs of sentences into one of three categories: Entailment, Contradiction, or Neutral. Unlike the original cross-lingual design of XNLI, we conducted monolingual experiments for each language to evaluate the performance of our models individually in each linguistic context.

Supplementary details regarding data preprocessing for downstream experiments are available in Appendix B.

We evaluate the performances of the encoder output representations for the 2-MT and 2-LM models and of the last hidden representation before the vocabulary projection for the single-stack models.

The evaluation of the models involves two distinct experimental approaches to test the performance: probing and fine-tuning. In the probing experiments, only the parameters of the classification heads are adjusted. This method primarily tests the raw capability of the pre-trained models' embeddings to adapt to specific tasks with minimal parameter changes, preserving the underlying pre-trained network structure. Conversely, in the fine-tuning experiments, all parameters of the models are adjusted. This approach allows the entire model to adapt to the specifics

of the task, potentially leading to higher performance at the cost of significantly altering the pre-trained weights.

For both experimental approaches, each model is trained for 10 epochs to ensure sufficient learning without overfitting. We optimize parameters with AdamW (Loshchilov and Hutter, 2017), with a constant learning rate of 0.0001 across all tasks and models. This setup was chosen to standardize the training process, providing a fair basis for comparing the performance outcomes across different models and tasks. We reproduce probing and fine-tuning for 5 seeds to ensure stability.

3 Results

3.1 Double-stack models

We first compare the performance of 2-LM and 2-MT across several key language processing tasks including SA, NER, POS tagging, and NLI. Results are shown in Tables 1a and 1b. The pretraining objectives play a significant role in shaping the models' effectiveness. Specifically, 2-MT, which is pretrained with a machine translation objective, consistently outperforms 2-LM, which utilizes a denoising objective. This pattern is consistent across all languages tested after fine-tuning as well as probing.

3.2 Single-stack models

Turning to the single-stack models (CLM, MLM, TLM), we find a somewhat more complex picture. In a probing context (cf. Table 2a), we find the CLM to be almost always the most effective, except for NLI in five languages and NER in Arabic, where it performs slightly less favorably compared to the MLM. As for fine-tuning (Table 2b), while the MLM generally ranks first on all POS, NER, and NLI datasets, the TLM is usually effective for SA.3

Table 1a, 1b, 2a, 2b and detailed results are mentioned but not provided in the text. [MISSING]

Figure 1: Performance tables for double-stack and single-stack models.

4 Related Work

ing in BART on cross-lingual downstream tasks. Monolingual evaluation of multilingual systems has also been broached a.o. by Rust et al. (2021).

5 Conclusion

This paper conducts an empirical study of how pretraining conditions of multilingual models impact downstream performances in probing and fine-tuning scenarios. Despite the inherent limitations that stem from our stringent data requirements, our experiments offer a novel perspective that highlights directions for future inquiry into how multilingual foundation models ought to be pretrained. We observe that double-stack BART-based models fare much better than single-stack models in probing scenarios, but the difference is overall less clear when it comes to fine-tuning. We also find some tentative evidence that translation objectives can be highly effective for model pretraining in precise circumstances: Namely, the most effective model on downstream tasks among those we experimented with is an MT-pretrained BART-like model,

which outperforms both a more traditional denoising objective for BART as well as decoder-only CLM and encoder-only MLM models. This would suggest that translation can serve as a powerful pretraining objective, although it is currently under-explored.4

Another crucial aspect of our study is that we present strictly comparable models, trained on comparable data, with comparable parameter counts and unified implementations. While this entails some limitations, especially with regard to the scale of models and data used, we nonetheless believe that a strict comparison can help discriminate between the various factors at play in other works. Here, we find clear evidence that CLM pretraining objectives, such as those used in GPT, outperform MLM-based models, such as BERT, in probing scenarios; we are also able to isolate and highlight how the optimal choice of pretraining objective is contingent on the architecture being employed.

text[[140, 751, 483, 849], [508, 72, 853, 121]] For future work, we recommend exploring multitask learning during pretraining by combining objectives like translation, denoising, and language modeling; in such cases, models could harness the strengths of each task to become more robust and versatile. Additionally, investigating training free evaluation methods can offer insights into a model's inherent capabilities without the variability introduced by fine-tuning.

Acknowledgments

We thank Alessandro Raganato and our colleagues at the Helsinki-NLP group for useful discussions throughout this project, as well as the three anonymous reviewers for their comments.

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546], and partially funded by the French National Research Agency [grant ANR-23-IAS1-0001]. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

The authors wish to thank CSC-IT Center for Science, Finland, for the generous computational resources on the Puhti supercomputer and LUMI supercomputer through the LUMI extreme scale access (MOOMIN and LumiNMT). Some of the experiments were performed using the Jean Zay and Adastra clusters from GENCI-IDRIS [grant 2022 A0131013801].

6 Limitations

This study employs models that are not large in terms of parameters in the era of large language models. Such a constraint potentially hinders the generalizability of our results to much larger architectures that are capable of handling a broader array of linguistic nuances. Furthermore, our study focuses on a small selected group of languages and specific NLP tasks. This focus might limit the applicability of our findings to other linguistic contexts or more complex real-world applications where diverse language phenomena or different task demands play a crucial role.

Another limitation is our reliance on specific corpora. The datasets utilized, while valuable, represent a potential source of selection bias. They may not fully encompass the vast diversity of global language use, thus skewing the model training and evaluation. Such a bias could affect the robustness and effectiveness of the pretrained models when applied to languages that are not well-represented in the training data.

References

References

[1] Devlin et al., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[2] Muennighoff et al., 2022. Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786.

[3] Alves et al., 2024. [MISSING].

[4] Liu et al., 2020. Multilingual denoising pretraining for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726-742.

[5] Xue et al., 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.

[6] Gorman and Bedrick, 2019. We need to talk about standard splits. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2786-2791.

[7] Wu and 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.

[8] Wang et al., 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.

[9] Cao et al., 2020. [MISSING].

[10] Wu and 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.

[11] Vaswani et al., 2017. Attention is all you need. In Advances in Neural Information Processing Systems.

[12] Ott et al., 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48-53.

[13] Lewis et al., 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.

[14] Radford et al., 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

[15] Conneau and Lample, 2019. [MISSING].

[16] Ziemski et al., 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.

[17] Tiedemann, 2012. Parallel data, tools and interfaces in opus. In Proceedings of LREC, volume 2012, pages 2214-2218.

[18] Sennrich et al., 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715-1725.

[19] Kingma and Ba, 2017. Adam: A method for stochastic optimization.

[20] Hou et al., 2024. Bridging language and items for retrieval and recommendation. arXiv preprint arXiv:2403.03952.

[21] Smetanin and Komarov, 2019. Sentiment analysis of product reviews in russian using convolutional neural networks. In 2019 IEEE 21st Conference on Business Informatics (CBI), volume 01, pages 482-486.

[22] ElSahar and El-Beltagy, 2015. [MISSING].

[23] Fetahu et al., 2023. Multi-CoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2027-2051.

[24] Malmasi et al., 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In Proceedings of the 29th International Conference on Computational Linguistics, pages 3798-3809.

[25] Mohit et al., 2012a. Recoll-oriented learning of named entities in Arabic Wikipedia. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 162-173.

[26] Nivre et al., 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. arXiv preprint arXiv:2004.10643.

[27] Conneau et al., 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475-2485.

[28] Williams et al., 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112-1122.

[29] Loshchilov and Hutter, 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

[30] Rust et al., 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3118-3135.

A Overview of pretraining objectives

Table 3 displays an example data point for all pretraining objectives we consider. In principle, the CLM is a document-level objective, i.e., the full document would be used as an input rather than the two sentences we show here.

B Datasets statistics

An overview of the volume of data available for pretraining is displayed in Table 4. The majority of the data were used for training.

In Table 5, we present an overview of the datasets used for downstream evaluation.

C Detailed results

In Table 6 and Table 7, we present the macro-f1 score of models in the downstream evaluation.

Table 1: Overview of the different objectives considered in this study. Top two rows: two-stacks (encoder-decoder) models; bottom three rows: single-stack (encoder-only or decoder-only) models. [ILLEGIBLE]

Table 2: Number of sentences in pretraining corpora.

	Train	Validation	Test	Total
UNPC	114,376,177	76,303	40,712	114,493,192
OpSub	81,622,353	359,035	77,342	82,058,730
Total	195,998,530	435,338	118,054	196,551,922

Table 3: Statistics of datasets used for downstream evaluation tasks.

Task	Language	Dataset
SA	EN	Amazon Review (Hou et al., 2024)
	ES	Amazon Review (Hou et al., 2024)
	FR	Amazon Review (Hou et al., 2024)
	ZH	Amazon Review (Hou et al., 2024)
	RU	RuReviews (Smetanin and Komarov, 2019)
	AR	ar_res_reviews (ElSahar and El-Beltagy, 2015)
NER	EN	MultiCoNER v2 (Fetahu et al., 2023)
	ES	MultiCoNER v2
	FR	MultiCoNER v2
	ZH	MultiCoNER v2
	RU	MultiCoNER v1 (Malmasi et al., 2022)
	AR	AQMAR Wikipedia NER corpus (Mohit et al., 2012b)
POS	EN	UD_English-GUM (Zeldes, 2017)
	ES	UD_Spanish-GSD (McDonald et al., 2013)
	FR	UD_French-GSD (Guillaume et al., 2019)
	ZH	UD_Chinese-Big (Nivre et al., 2017)+UD_Chinese-HK (Wong et al., 2017)+UD_Chinese-CFL (Lee et al.,
	RU	UD_Russian-Taiga (Lyashevskaya et al., 2018)
	AR	UD_Arabic-PADT (Zemánek, 2008)
NLI	EN	XNLI (Conneau et al., 2018)
	ES	XNLI
	FR	XNLI
	ZH	XNLI
	RU	XNLI
	AR	XNLI

Table 4: Macro F1 score after model fine-tuning. [MISSING]