2/9/2026 4:04:32 PM

Compare Results

Old File:		New File:
2024.emnlp-main.888.pdf	versus	**2024_emnlp-main_888.pdf**
13 pages (220 KB)		**11 pages (110 KB)**
10/31/2024 10:33:02 PM		2/9/2026 2:13:51 PM

Total Changes

874

Content

82 Replacements
113 Insertions
287 Deletions

Styling and Annotations

156 Styling
236 Annotations

Go to First Change (page 2)

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@{1helsinki.fi, 2univ-ubs.fr}

February 9, 2026

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 (?) 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., ?; ?). There is considerable variation as to how such models are trained: For instance, some rely on datasets comprising multiple languages without explicit cross-lingual supervision (e.g., ?), and some use explicit supervision (?).

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 (?). 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:

*Equal contribution and corresponding authors.

[label=()]

1. Does the explicit cross-lingual training signal of translation objectives foster better downstream performances in monolingual tasks?

2. 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 (e.g., **?**), yet explicit alignments such as linear mapping (**?**) and L2 alignment (**?**) between source and target languages do not necessarily improve the quality of cross-lingual representations (**?**).

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 (**?**) and implemented in fairseq[1] (**?**). 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 (**?**). 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.

We also consider three single-stack models: (i) an encoder-only model trained on the masked language modeling objective (MLM) (**?**); (ii) an autoregressive causal language model (CLM), similar to GPT, 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 (**?**). We provide an example datapoint for each pretraining objective in Table **??**, Appendix **??**.

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 (**?**) and OpenSubtitles (OpSub; **?**) 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

[1] In this work, we only focus on the causal variant of TLM proposed by Conneau and Lample.

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??, Appendix ??.

To ensure a fair comparison, we control key variables, including tokenization (100k BPE pieces; ?), number of transformer layers (12), hidden dimensions (512), attention heads (8), and feedforward layer dimensions (2048). We perform 600k steps of updates,[2] 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 optimiz(??). 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 datas(??) in English, Spanish, French, and Chinese. RuRevie(?) for Russian, and ar_res_revie(?) 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 datas(?) for English, Spanish, French, and Chinese, MultiCoNER v1(?) for Russian and the AQMAR Wikipedia NER corpus(?) 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 datas(), 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 datas(?) for the six languages. The XNLI dataset consists of sentence pairs translated from the MultiNLI datas(?) 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 Append(??.

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(?), 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

[2] Improvements in cross-entropy over the validation set were always marginal after this stage.

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 **??** and **??**. The pretraining objectives play a significant role in shaping the models' effectiveness. Specifically, 2-MT, which is pre-trained 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.

Table 1: Accuracy (×100) of double-stack models (± s.d. over 5 runs) – Probing

Setup	Languages	EN	ES	FR	ZH	RU	AR
SA	2-LM	42.86±0.86	42.80±0.69	43.00±0.60	40.41±1.02	65.83±0.70	70.88±1.62
	2-MT	46.71±0.88	46.64±0.55	46.10±0.43	43.74±0.65	68.79±0.42	73.77±0.97
NER	2-LM	82.69±0.09	84.74±0.07	82.80±0.06	78.88±0.25	77.93±0.15	85.28±0.22
	2-MT	89.47±0.06	90.54±0.04	89.41±0.10	88.78±0.09	83.39±0.22	89.70±0.18
POS	2-LM	78.85±0.29	78.12±0.25	81.57±0.32	66.09±0.25	77.93±0.12	47.68±0.10
	2-MT	92.22±0.14	90.59±0.20	95.39±0.10	75.87±0.17	93.20±0.08	61.84±0.24
NLI	2-LM	48.56±0.01	49.31±0.01	48.33±0.01	38.81±0.01	48.34±0.01	45.11±0.01
	2-MT	60.50±0.01	59.56±0.01	59.00±0.01	59.01±0.01	59.83±0.01	59.58±0.01

Table 2: Accuracy (×100) of double-stack models (± s.d. over 5 runs) – Fine-tuning

Setup	Languages	EN	ES	FR	ZH	RU	AR
SA	2-LM	52.26±0.55	52.89±0.69	52.99±0.59	48.64±0.36	73.89±0.43	79.74±1.36
	2-MT	54.76±0.58	55.56±0.49	54.75±0.42	50.55±0.68	74.77±0.50	81.49±1.49
NER	2-LM	91.13±0.12	91.82±0.21	91.58±0.10	92.30±0.10	85.34±0.39	89.05±0.13
	2-MT	93.46±0.09	94.22±0.09	93.84±0.04	93.75±0.32	89.07±0.11	93.26±0.15
POS	2-LM	92.42±0.28	90.41±0.16	95.21±0.13	82.30±0.48	95.36±0.20	69.57±0.24
	2-MT	95.98±0.08	94.29±0.05	98.05±0.17	90.18±0.15	97.00±0.07	74.47±0.08
NLI	2-LM	57.76±0.01	57.87±0.01	56.77±0.01	48.05±0.01	56.43±0.01	53.77±0.01
	2-MT	61.96±0.01	61.71±0.01	60.09±0.01	53.72±0.01	59.00±0.01	56.93±0.01

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 **??**), 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 **??**), while the MLM generally ranks first on all POS, NER, and NLI datasets, the TLM is usually effective for SA.[3]

[3]However, remark that unlike with the BART-based models, SA results are not stable when we shift metrics from accuracy to F1 (see Tables **??** and **??** in Appendix **??**). The difference in F1 between the top two models is often ≤ 0.01, making it difficult to ascertain that one model strictly dominates.

Table 3: Accuracy (×100) of single-stack models (± s.d. over 5 runs) – Probing

Setup	Languages	EN	ES	FR	ZH	RU	AR
SA	CLM	35.14±0.92	35.66±1.10	34.14±1.63	33.62±0.83	57.57±1.11	67.71±2.24
	MLM	34.26±1.34	34.82±1.58	33.90±1.12	32.52±1.65	54.55±1.86	65.94±3
	TLM	29.68±2.22	32.20±3.07	32.26±2.34	29.88±4.17	56.45±1.81	64.45±1.81
NER	CLM	80.27±0.12	82.59±0.06	80.38±0.12	77.92±0.28	76.39±0.03	84.17±0.08
	MLM	78.77±0.02	81.61±0.00	79.11±0.01	70.67±0.10	76.34±0.01	84.29±0.00
	TLM	79.10±0.06	81.94±0.13	79.56±0.14	77.26±0.24	76.39±0.02	84.26±0.02
POS	CLM	69.06±0.38	70.32±0.50	76.67±0.46	51.40±0.47	59.64±0.62	43.49±0.40
	MLM	37.92±0.61	44.26±0.11	46.89±0.32	31.16±0.21	34.62±0.16	34.71±0.94
	TLM	62.96±1.02	62.08±1.99	63.89±1.06	50.46±0.53	54.27±0.87	40.94±1.16
NLI	CLM	42.32±0.02	42.99±0.01	43.43±0.02	40.55±0.02	40.06±0.02	41.99±0.01
	MLM	45.64±0.02	44.49±0.01	43.11±0.02	42.80±0.01	43.16±0.01	43.55±0.01
	TLM	38.36±0.02	41.95±0.02	41.89±0.01	38.93±0.04	41.20±0.02	39.50±0.02

Table 4: Accuracy (×100) of single-stack models (± s.d. over 5 runs) – Fine-tuning

Setup	Languages	EN	ES	FR	ZH	RU	AR
SA	CLM	55.23±0.72	47.81±15.55	54.84±0.62	51.18±0.94	75.07±0.21	66.18±21.74
	MLM	55.22±0.92	55.67±1.77	54.08±2.43	51.00±1.07	74.53±1.36	75.00±3.48
	TLM	55.14±0.92	55.84±0.59	55.22±0.98	51.46±0.53	75.31±0.57	72.75±2.25
NER	CLM	89.91±0.33	91.42±0.15	90.65±0.17	89.97±0.14	83.20±0.31	87.50±2.22
	MLM	93.31±0.57	93.93±0.60	93.67±0.30	92.99±0.99	87.49±0.78	85.78±3.30
	TLM	89.88±0.06	91.45±0.25	90.49±0.23	90.10±0.14	83.76±0.63	84.29±0.00
POS	CLM	91.72±0.14	90.51±0.13	95.75±0.10	78.61±0.31	85.50±0.15	57.43±1.63
	MLM	96.00±0.15	94.45±0.13	97.94±0.20	89.96±0.71	96.69±0.13	74.35±0.53
	TLM	91.68±0.19	90.38±0.20	86.99±19.40	78.50±0.52	85.71±0.18	59.11±0.50
NLI	CLM	48.84±0.14	56.46±0.03	55.45±0.03	49.70±0.06	55.23±0.02	49.02±0.07
	MLM	59.41±0.01	57.54±0.04	55.04±0.06	47.96±0.03	57.80±0.01	53.60±0.01
	TLM	49.76±0.10	52.12±0.11	54.20±0.10	49.03±0.04	53.60±0.04	44.39±0.10

3.3 Discussion

A first global observation that we can make for these results is that single-stack and double-stack models appear to behave differently. While the MT objective yields the highest performances for BART-type models, the downstream performances of the TLM do not really stand out compared to the CLM in probing and the MLM in fine-tuning scenarios. It is important to note that the performances stem at least in part from the architecture itself: 2-MT and 2-LM both consistently outperform all single-stack models in probing. However, it is crucial to acknowledge the limitations of our study, as we only conducted one pretraining round for all the objectives. Hence, this evidence should be interpreted as tentative at best.

Fine-tuning also tends to minimize the difference between single-stack and double-stack models—which suggests that the higher quality of double-stack representations could be an artifact of training limitations. Moreover, the relative ranks of the three single-stack models fluctuate much more than what we see for the double-stack models, owing to no little extent to the oftentimes momentous variation across seeds for single-stack models. We therefore conjecture that while a translation objective can yield a clear training signal towards semantically informed representations, this comes with two caveats:

first, the signal can only be leveraged with dedicated separate modeling of source and target (viz. double-stack models); second, this advantage is much less consequential when fine-tuning.

4 Related works

Multilingual foundation models have flourished in recent years (e.g., ?????????, and with them so have studies of their representations (e.g., ?????). All of these works, however, fail to control for some of the most crucial factors, such as ensuring that all models are trained on comparable amounts of data.

This work is specifically related to ?, which also compares MLM, CLM, and TLM but does not normalize the training data. Another point of comparison is ?, which studies the impact of MT continued pretraining in BART on cross-lingual downstream tasks. Monolingual evaluation of multilingual systems has also been broached e.g. by ?.

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.

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

[4] There are reasonable objections against using MT models as pretrained multilingual foundation models—namely, unlike auto-regressive causal language models, their generation capabilities are strictly tied to translation, thereby requiring some degree of multilingualism from end-users.

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].

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.

A Overview of pretraining objectives

Table ?? 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 ??. The majority of the data were used for training.

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

C Detailed results

In Tables ?? and ??, we present the macro-F1 score of models in the downstream evaluation.

Table 5: 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.

Objective	Source input	Target output
2-LM	_D'autres_mesures_de_ce_type_vont_être[MASK][MASK],_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de[MASK]_du_voyage_(« Camminanti »).</s>	<s>_D'autres_mesures_de_ce_type_vont_être_appliquées,_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de_gens_du_voyage_(« Camminanti »).</s>
2-MT	<fr>_D'autres_mesures_de_ce_type_vont_être_appliquées,_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de_gens_du_voyage_(« Camminanti »).	<s>_Other_similar_measures_are_going_to_be_taken_in_cooperation_with_other_Roma,_Sinti_and_Travellers_("Camminanti")_associations.</s>
CLM	..._Divers_accords_ad_hoc_ont_été_conclus_à_cet_effet_par_le_Ministère_de_l'éducation_et_l'association_Nomadi._Des_mesures_de_ce_type_vont_être_appliquées,_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de_gens_du_voyage_(« Camminanti »)....	..._accords_ad_hoc_ont_été_conclus_à_cet_effet_par_le_Ministère_de_l'éducation_et_l'association_Nomadi._Des_mesures_de_ce_type_vont_être_appliquées,_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de_gens_du_voyage_(« Camminanti »)....
TLM	_D'autres_mesures_de_ce_type_vont_être_appliquées,_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de_gens_du_voyage_(« Camminanti »).<fr2en>_Other_similar_measures_are_going_to_be_taken_in_cooperation_with_other_Roma,_Sinti_and_Travellers_("Camminanti")_associations._	_mesures_de_ce_type_vont_être_appliquées,_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de_gens_du_voyage_(« Camminanti »).<fr2en>_Other_similar_measures_are_going_to_be_taken_in_cooperation_with_other_Roma,_Sinti_and_Travellers_("Camminanti")_associations. </s>
MLM	<s>_D'autres_mesures_de_ce_type_vont_être[MASK][MASK],_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de[MASK]_du_voyage_(« Camminanti »).</s>	<s>_D'autres_mesures_de_ce_type_vont_être_appliquées,_en_coopération_avec_d'autres_associations_de_Roms,_de_Sintis_et_de_gens_du_voyage_(« Camminanti »).</s>

Table 6: 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 7: Statistics of datasets used for downstream evaluation tasks.

Task	Language	Dataset	Class Count	Train	Validation	Test	Total
SA	EN	Amazon Review	5	200,000	5,000	5,000	210,000
	ES	Amazon Review	5	200,000	5,000	5,000	210,000
	FR	Amazon Review	5	200,000	5,000	5,000	210,000
	ZH	Amazon Review	5	200,000	5,000	5,000	210,000
	RU	RuReviews	3	85,601	2,143	2,137	89,881
	AR	ar_res_reviews	2	6,680	835	835	8,350
NER	EN	MultiCoNER v2	3	253,011	13,323	3,773,671	4,040,005
	ES	MultiCoNER v2	3	262,814	13,462	3,925,900	4,202,176
	FR	MultiCoNER v2	3	247,743	13,062	3,742,924	4,003,729
	ZH	MultiCoNER v2	3	245,606	12,816	489,605	748,027
	RU	MultiCoNER v1	3	242,384	12,787	2,061,318	2,316,489
	AR	AQMAR Wikipedia NER	3	57,053	8,615	8,185	73,853
POS	EN	UD_English-GUM	16	128,391	16,070	15,554	160,015
	ES	UD_Spanish-GSD	16	127,459	16,916	15,645	160,020
	FR	UD_French-GSD	15	127,638	16,207	16,167	160,012
	ZH	Multiple UD treebanks	16	128,935	15,680	15,758	160,373
	RU	UD_Russian-Taiga	16	127,647	16,175	16,184	160,006
	AR	UD_Arabic-PADT	16	127,552	16,608	15,848	160,008
NLI	EN	XNLI	3	392,702	2,490	5,010	400,202
	ES	XNLI	3	392,702	2,490	5,010	400,202
	FR	XNLI	3	392,702	2,490	5,010	400,202
	ZH	XNLI	3	392,702	2,490	5,010	400,202
	RU	XNLI	3	392,702	2,490	5,010	400,202
	AR	XNLI	3	392,702	2,490	5,010	400,202

Table 8: Macro F1 score using probing technique.

Task	Model	EN	ES	FR	ZH	RU	AR
SA	2-LM	0.4130±0.0118	0.4120±0.0160	0.4166±0.0076	0.3859±0.0156	0.6599±0.0101	0.6343±0
	2-MT	0.4588±0.0092	0.4554±0.0053	0.4448±0.0158	0.4260±0.0070	0.6935±0.0052	0.6864±0
	CLM	0.3183±0.0099	0.3351±0.0198	0.3066±0.0192	0.3104±0.0135	0.5693±0.0107	0.5886±0
	MLM	0.3236±0.0270	0.3188±0.0188	0.3153±0.0088	0.2936±0.0107	0.5434±0.0236	0.5804±0
	TLM	0.2593±0.0298	0.2768±0.0589	0.2528±0.0487	0.2344±0.0539	0.5537±0.0307	0.5487±0
NER	2-LM	0.5830±0.0057	0.5616±0.0070	0.5627±0.0039	0.5653±0.0164	0.4178±0.0100	0.4310±0
	2-MT	0.7778±0.0014	0.7660±0.0014	0.7716±0.0031	0.7871±0.0043	0.6551±0.0088	0.7311±0
	CLM	0.4516±0.0110	0.4213±0.0075	0.4306±0.0131	0.5086±0.0053	0.3004±0.0034	0.3223±0
	MLM	0.3003±0.0017	0.2997±0.0001	0.3021±0.0019	0.3341±0.0108	0.2891±0.0001	0.3094±0
	TLM	0.3485±0.0074	0.3471±0.0152	0.3499±0.0173	0.4876±0.0230	0.2941±0.0015	0.3094±0
POS	2-LM	0.7241±0.0040	0.6607±0.0042	0.6848±0.0074	0.5964±0.0072	0.7427±0.0030	0.4678±0
	2-MT	0.8520±0.0065	0.7685±0.0203	0.8300±0.0017	0.7002±0.0029	0.8587±0.0055	0.6575±0
	CLM	0.5621±0.0069	0.5422±0.0066	0.5568±0.0064	0.3761±0.0148	0.4975±0.0140	0.3040±0
	MLM	0.2157±0.0063	0.1499±0.0055	0.1722±0.0084	0.0717±0.0040	0.1275±0.0080	0.1511±0
	TLM	0.4741±0.0147	0.3759±0.0378	0.3744±0.0153	0.3314±0.0112	0.3798±0.0097	0.2299±0
NLI	2-LM	0.4825±0.0075	0.4901±0.0046	0.4779±0.0102	0.3805±0.0089	0.4804±0.0059	0.4445±0
	2-MT	0.6017±0.0105	0.5938±0.0119	0.5860±0.0087	0.5881±0.0031	0.5982±0.0025	0.5943±0
	CLM	0.3946±0.0479	0.4134±0.0227	0.4068±0.0373	0.3744±0.0400	0.3593±0.0519	0.3978±0
	MLM	0.4464±0.0328	0.4330±0.0145	0.4157±0.0347	0.4208±0.0110	0.4162±0.0251	0.4281±0
	TLM	0.3063±0.0361	0.3573±0.0327	0.3940±0.0240	0.3122±0.0876	0.3892±0.0390	0.3360±0

Table 9: Macro F1 score after model fine-tuning.

Task	Model	EN	ES	FR	ZH	RU	AR
SA	2-LM	0.5213±0.0068	0.5254±0.0083	0.5244±0.0135	0.4739±0.0096	0.7421±0.0059	0.7522±0
	2-MT	0.5407±0.0086	0.5510±0.0084	0.5398±0.0054	0.4956±0.0093	0.7522±0.0056	0.7767±0
	CLM	0.5443±0.0072	0.4446±0.2115	0.5421±0.0089	0.5015±0.0187	0.7553±0.0015	0.5283±0
	MLM	0.5441±0.0107	0.5466±0.0314	0.5348±0.0237	0.4972±0.0142	0.7509±0.0135	0.5695±0
	TLM	0.5358±0.0186	0.5501±0.0128	0.5474±0.0137	0.5069±0.0119	0.7586±0.0057	0.4599±0
NER	2-LM	0.8200±0.0042	0.8092±0.0053	0.8259±0.0035	0.8626±0.0022	0.7215±0.0122	0.7274±0
	2-MT	0.8670±0.0017	0.8651±0.0022	0.8727±0.0018	0.8897±0.0042	0.7934±0.0039	0.8685±0
	CLM	0.7950±0.0064	0.8053±0.0028	0.8099±0.0044	0.8129±0.0021	0.6622±0.0182	0.5994±0
	MLM	0.8635±0.0123	0.8580±0.0142	0.8706±0.0055	0.8739±0.0199	0.7629±0.0172	0.4113±0
	TLM	0.7908±0.0028	0.8024±0.0081	0.8067±0.0047	0.8120±0.0032	0.6758±0.0312	0.3094±0
POS	2-LM	0.8925±0.0039	0.7365±0.0025	0.8496±0.0034	0.8088±0.0059	0.8984±0.0055	0.7769±0
	2-MT	0.9314±0.0024	0.7826±0.0235	0.8866±0.0074	0.8842±0.0059	0.9285±0.0029	0.8660±0
	CLM	0.8752±0.0042	0.7854±0.0024	0.8573±0.0041	0.7906±0.0195	0.8264±0.0104	0.5932±0
	MLM	0.9177±0.0068	0.8079±0.0259	0.8851±0.0019	0.8313±0.0079	0.9226±0.0048	0.8602±0
	TLM	0.8782±0.0045	0.7830±0.0067	0.7421±0.2503	0.7876±0.0271	0.8247±0.0088	0.6201±0
NLI	2-LM	0.5771±0.0067	0.5760±0.0088	0.5658±0.0085	0.4766±0.0058	0.5629±0.0052	0.5350±0
	2-MT	0.6183±0.0054	0.6151±0.0082	0.5991±0.0073	0.5302±0.0086	0.5887±0.0041	0.5678±0
	CLM	0.4240±0.2315	0.5589±0.0355	0.5493±0.0404	0.4729±0.1123	0.5507±0.0265	0.4554±0
	MLM	0.5927±0.0189	0.5719±0.0487	0.5282±0.0964	0.4618±0.0453	0.5775±0.0069	0.5247±0
	TLM	0.4428±0.1751	0.4728±0.1731	0.5345±0.1076	0.4558±0.0722	0.5061±0.0771	0.3816±0