Not Enough Data to Pre-train Your Language Model? MT to the Rescue!

Gorka Urbizu[12] **Iñaki San Vicente**[1] **Xabier Saralegi**[1] **Ander Corral**[1]

[1] **Orai NLP Technologies**
[2] University of the Basque Country
[g.urbizu,i.sanvicente,x.saralegi,a.corral]@orai.eus

Abstract

In recent years, pre-trained transformer-based language models (LM) have become a key resource for implementing most NLP tasks. However, pre-training such models demands large text collections not available in most languages. In this paper, we study the use of machine-translated corpora for pre-training LMs. We answer the following research questions: RQ1: Is MT-based data an alternative to real data for learning a LM?; RQ2: Can real data be complemented with translated data and improve the resulting LM? In order to validate these two questions, several BERT models for Basque have been trained, combining real data and synthetic data translated from Spanish. The evaluation carried out on 9 NLU tasks indicates that models trained exclusively on translated data offer competitive results. Furthermore, models trained with real data can be improved with synthetic data, although further research is needed on the matter.

1 Introduction

Since the emergence of the attention-based Transformer architecture (Vaswani et al., 2017) and the masking pre-training strategies introduced by BERT (Devlin et al., 2019), transformer-based language models have become the default approach for many NLP tasks, leading to an impressive performance in high-resource languages, particularly English (Hoffmann et al., 2022; Thoppilan et al., 2022; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022).

As scaling laws dictate (Kaplan et al., 2020; Hoffmann et al., 2022), such competitive models are achievable with big computational budgets and large corpora available, requirements difficult to meet for most languages (Joshi et al., 2020).

Fortunately, LMs are being built for less-resourced languages, such as KinyaBERT for Kinyarwanda (390M words) (Nzeyimana and Rubungo, 2022), ElhBERTeu for Basque (351M) (Urbizu et al., 2022), gaBERT for Irish (161M) (Barry et al., 2021), LuxemBERT for Luxembourgish (130M) (Lothritz et al., 2022b), Bertinho 45M for Galician (Vilares et al., 2021), swahBERT for Swahili (16M) (Martin et al., 2022) and QuBERT for Quechua (4M) (Zevallos et al., 2022).

Zhang et al. (2021) estimates that 10M-100M words of pre-training data are enough for an LM to acquire the linguistic capacities of syntax and semantics, but the amount of data required to acquire factual knowledge and commonsense is higher.

In this work, we propose to tackle the lack of data by using text corpora available in other languages translated via Machine Translation (MT). To the best of our knowledge, this has been addressed before (Lothritz et al., 2022a) but not in-depth, and only for a closely related language pair (German-Luxembourgish). We selected Basque, a language isolate, as the target language, and employ Spanish as the auxiliary language.

We direct our efforts to answer the following Research Questions (RQ):

RQ1: *Can we obtain comparable performance to a native LM by training LMs just on synthetic data from MT*?

RQ2: *Can we improve current LMs for less-resourced languages by adding synthetic MT data*?

2 Methodology

In order to answer our research questions, we set out the following methodology. We propose two baseline LMs: i) ElhBERTeu (Urbizu et al., 2022) as a strong baseline, trained on a corpus of 351M words; and (ii) BERT_{125M} a model trained on a lower data regime. From there on we pre-train various LMs with different native/synthetic data combinations. Sections 3 and 4 give details of the models pre-trained, including baselines. All models presented in this paper follow the BERT base architecture (Devlin et al., 2019).

We select Basque, a language isolate, as a tar-

Findings of the Association for Computational Linguistics: ACL 2023, pages 3826–3836
July 9-14, 2023

get language, and employ Spanish, a Romance language, as the auxiliary language since it has huge text corpora available. Furthermore, both languages coexist in the same geographical area, therefore, Spanish is the language that Basque shares the most parallel data with, which is crucial to train MT systems. On the other hand, this is a real case since obtaining a corpus in Basque that exceeds 350M words is difficult.

2.1 MT system

The Spanish to Basque MT system used for our experiments is based on the default Base Transformer architecture (Vaswani et al., 2017) using the PyTorch version of the OpenNMT toolkit (Klein et al., 2017) and BPE tokenization (Sennrich et al., 2016) (joint vocabulary of 32K). The system was trained with 8.6M parallel sentences and evaluated on the FLORES-200 benchmark (Team et al., 2022) obtaining 13.2 BLEU and 47.4 chrF++. See Appendix F for an analysis of the impact the amount of parallel data has.

2.2 Corpora

Following we introduce the corpora employed on the experiments (summarized in Table 1):

N_ElhBERTeu is a Basque corpus compiled to train ElhBERTeu (Urbizu et al., 2022). It contains 351M words.

N_small is a smaller Basque native corpus (125M words), created to be closer to the scenario of many languages. The corpus is composed of 75% news articles from Berria[1] newspaper and 25% of text from Wikipedia.

S_beto2eu is the *Spanish Unannotated Corpora*[2] composed of 3B words (Cañete, 2019) which was used to train the Spanish LM BETO (Cañete et al., 2020), translated to Basque using the MT system described in Section 2.1.

S_loc2eu was also translated from Spanish to Basque. We collected up to 548M words of news articles in Spanish from news sources geographically limited to the Basque Country. After translating it with our MT system, the final corpus in Basque contains 378M words.

2.3 Pre-training Details

Since the aim of this work focuses on the effect of the training data, we left all the hyper-papameters fixed. Every model was pre-trained following the procedure used for ElhBERTeu (Urbizu et al., 2022). See appendix A for further details.

Corpora	Words	Domain
N_ElhBERTeu	351M	Mix
N_small	125M	News+Wiki
S_beto2eu	2.17B	Mix
S_loc2eu	378M	News

Table 1: Corpora used to train our models. Word count for synthetic text is done once translated to Basque.

2.4 Evaluation

A downstream task evaluation of our models was performed on the BasqueGLUE (Urbizu et al., 2022) NLU benchmark for Basque. BasqueGLUE includes the following tasks: Name Entity recognition (NERC), Intent Classification (intent)(de Lacalle et al., 2020), Slot Filling (slot)(de Lacalle et al., 2020), Topic Classification (BHTC)(Agerri et al., 2020), Sentiment Analysis (BEC), Stance Detection (Vaxx)(Agerri et al., 2021), QA-NLI (QNLI), Word in Context (WiC) and Coreference Resolution (coref). Metrics employed are accuracy in QNLI, WiC and coref, Macro F1-score in Vaxx, and Micro F1-score in the remaining tasks.

We fine-tuned each model up to 10 epochs and selected the optimal number of epochs over the development set. We use a batch size of 32 and a learning rate of 3e-5. We report the average of 5 runs on the test sets. Fine-tuning was done on NVIDIA GeForce RTX 3090 GPUs.

3 LM Trained Solely on Synthetic Data

RQ1 aims to prove if it is possible to train a competitive LM with just synthetic text obtained from MT. In order to do that we train a BERT model on S_beto2eu (S_BERT), and evaluate if the model trained exclusively on synthetic data is able to perform as well as models trained on real data.

The results on the BasqueGLUE Benchmark for S_BERT are reported in Table 2[3]. S_BERT achieves competitive results. Although it does not perform as well as our strongest baseline ElhBERTeu, S_BERT, trained solely on translated texts, is comparable to BERT_{125M} (trained on N_small Basque native corpus), performing better depending on the task.

[1] www.berria.eus

[2] www.github.com/josecannete/spanish-corpora

[3] A version of this table with standard deviations and MLM task is available in Appendix D and Appendix E.

	avg	NERC	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref
ElhBERTeu	**73.40**	82.03	74.13	82.19	**78.48**	69.46	**76.04**	59.41	**72.27**	**66.64**
BERT$_{125M}$	71.98	79.51	**75.18**	80.83	76.94	70.15	73.76	58.32	70.09	63.10
S_BERT	71.40	79.72	74.03	80.94	73.32	68.83	73.93	58.92	70.06	62.83
SN_BERT	72.38	**82.33**	74.12	81.45	77.24	**70.32**	73.76	56.31	71.37	64.50
Sloc_BERT	72.21	79.74	74.75	**82.26**	75.91	69.80	72.66	**61.08**	70.37	63.27
SNloc_BERT	72.49	81.81	75.14	80.33	78.05	69.95	72.74	56.23	71.63	66.51

Table 2: Results for S_BERT (synthetic), SN_BERT (native+synthetic), Sloc_BERT (synthetic local) and SNloc_BERT (native+synthetic local) , compared with the native ElhBERTeu and BERT$_{125M}$ models.

	avg	NERC	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref
paral$_{EU}$	**69.19**	**73.88**	**74.04**	**81.82**	**72.56**	**68.88**	69.87	53.70	**68.67**	**59.28**
paral$_{ES2EU}$	68.65	72.42	72.60	78.12	71.09	68.10	**70.13**	**58.74**	68.13	58.50

Table 3: Downstream performance of the model trained on a native corpus (paral$_{EU}$ -29M-) vs. the model trained on the translated corpus of the same source (paral$_{ES2EU}$ -28M-).

In order to improve the results obtained with the synthetic data, we analyse two specific aspects of the data: i) the quality loss during the translation process; ii) the cultural context of the synthetic data. Following we analyze each of those factors.

3.1 Measuring the Quality of MT Text

To measure quality loss when translating from Spanish to Basque, we did a manual analysis on a sample of the translations produced by the MT system (See appendix B for details). We evaluated whether a sentence was correctly translated (71%), but also whether the produced sentence was linguistically correct (91%). Since we aim to use this text to train LMs, the effect of some translation errors like hallucinations or omissions, that cause significant meaning changes, might not be critical.

Next, we measured the vocabulary diversity loss during translation. For that aim, we compiled a Basque-Spanish parallel corpus (more details can be found in appendix B) and translated the Spanish text to Basque with our MT system. The lexicon of the translated data is 16% poorer, limited by the target vocabulary of the MT model and the tendency of MT to generalize and simplify the vocabulary.

Finally, we analyze the impact of training LMs on translated corpora, leaving aside other factors such as corpus size or text-domain. We train two BERT models using the parallel corpus compiled in the previous experiment, one on the original Basque part of the corpus and the other on the part translated from Spanish to Basque. The results in Table 3 show the model trained on translated data performs slightly worse than the native model. While this is expected from the quality loss and lexicon impoverishment caused by MT, the gap in performance is very small (0.5% on average), which leads to the conclusion that the synthetic data is adequate.

3.2 Domain and Cultural Context

Another factor related to data which might affect the performance of MLs trained over translated corpora is the source text we select in the auxiliary language. The Spanish Unannotated Corpora is a huge corpus. However, it is not domain homogeneous and the topic distribution of this corpus differs significantly from that of a corpus in Basque, especially because it hardly includes the specific topics associated with the Basque Country. Furthermore, we analyzed how tokenizers trained on this corpus do not include many words common in the context of Basque speaker communities, like named entities (locations, people or organizations). See appendix C for a detailed analysis of the vocabulary coverage of each model on the test datasets.

To analyze the impact the cultural bias and the domain heterogeneity of the source text has on the performance in downstream tasks, we compiled the S_loc2eu corpus, presented in section 2.2. This corpus is formed by texts in Spanish crawled from newspapers geographically and culturally connected to the Basque Country. Results in Table 2 show that models trained on translated local news (Sloc_BERT and SNloc_BERT), perform better than those without them (S_BERT and SN_BERT), even though it is trained over a much smaller corpus. Following the same pattern, the

	avg	NERC	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref
ElhBERTeu	73.40	82.03	74.13	**82.19**	78.48	69.46	**76.04**	59.41	**72.27**	66.64
concat_{20-80}	72.38	82.33	74.12	81.45	77.24	**70.32**	73.76	56.31	71.37	64.50
concat_{50-50}	73.12	81.99	**75.29**	79.87	77.80	69.23	72.41	**62.80**	71.93	66.78
concat_{80-20}	**73.47**	**82.05**	74.21	81.54	**78.57**	68.86	74.09	62.17	71.60	**68.11**
sequential	72.75	81.71	74.39	81.55	77.95	69.06	74.09	57.66	72.11	66.24

Table 4: Results of the models trained using different data combination strategies.

vocabulary coverage is higher for those models containing translated local news, as shown in appendix C.

4 Combining Native and Synthetic Data

The objective of RQ2 is to test if adding texts translated by MT to a native corpus can boost the performance on downstream NLU tasks of the LM in the target language.

With that aim, we trained a new LM on the concatenation of S_beto2eu corpus and the N_ElhBERTeu corpus[4] (SN_BERT hereinafter). Table 2 reports the results for SN_BERT when evaluated on the BasqueGLUE benchmark. Even if SN_BERT surpasses ElhBERTeu in a few tasks (NERC, BEC), it is below it in the average score.

4.1 Merging Strategies

One factor that may explain the lower performance of the model trained on the combined synthetic and native data is the way of combining the data. Our last experiment aims to analyze different combination alternatives. For SN_BERT, we just concatenate N_ElhBERTeu and S_beto2eu. However, the better quality native corpus is diluted among the translated texts of poorer quality, but larger in size (4x times). Hence, we propose another three alternatives to merge native and translated corpora, shifting the balance between both types of data:

concat_{20-80} (SN_BERT): concatenation of N_ElhBERTeu and S_beto2eu, which roughly form 20% and 80% of the pre-training corpus respectively. As mentioned, synthetic data take the principal role in this configuration.

concat_{50-50}: we oversample N_ElhBERTeu corpus to equal the size of S_beto2eu. This setting gives equal weight to native and synthetic data.

concat_{80-20}: we oversample N_ElhBERTeu up to 80%, thus, pre-training relies on native data mostly. Native data is weighted over synthetic data.

sequential: the LM is trained for 750K steps on S_beto2eu, and afterwards for another 250K steps on N_ElhBERTeu[5].

Results for different merging strategies are shown in Table 4. Increasing the ratio of N_ElhBERTeu data in our pre-training corpora improves the performance of our models to the point where concat_{80-20} outperforms ElhBERTeu, trained only with native text in Basque. Pre-training sequentially does improve slightly the results of the default SN_BERT setting, but weighting concatenation is the best strategy between the two. Further sequential training regimes were tried other than (750k+250K). 'sequential' refers to the best results we achieved with this strategy.

5 Conclusions

Regarding the RQ1, we conclude from our experiments that LMs trained exclusively on synthetic data from MT can obtain comparable performance to a native LM. We further analyze that other than the quality of MT, the cultural context of the text we select from the auxiliary language do have an effect on the final performance. We conclude that it is better to gather a corpus composed of sources similar to those in the target language, rather than indiscriminately translating vast amounts of data in the auxiliary language.

Furthermore, with respect to RQ2, our experiments show that state-of-the-art models' performance can be improved by adding translated data during the pre-training, albeit it is a small improvement. Weighting the native data above synthetic data is key to this improvement.

All in all, this approach has a big potential for less-resourced languages, since once you have a proper MT system, there is no limit on the amount of data one can translate from languages with bigger corpora available.

Data and models are publicly available[6].

[4]Concatenation is shuffled at document-level.

[5]Both phases use the tokenizer from N_ElhBERTeu

[6]https://github.com/orai-nlp/mt-bert

Limitations

In this work, we study the approach of using machine-translated text to train language models on a single target language (and language pair). Our conclusions may differ for other languages.

We use a fixed set of hyper-parameters during pre-training and only the epoch number is optimized during fine-tuning. Since our focus is the training-data used, and our resources are limited, we did not perform an extensive hyper-parameter search.

Acknowledgements

This work has been partially funded by the Basque Government (ICL4LANG project, grant no. KK-2023/00094). We also acknowledge the support of Google's TFRC program.

References

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Alvaro Rodrigo. 2021. Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788.

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J Ó Meachair, and Jennifer Foster. 2021. gabert–an irish language model. *arXiv preprint arXiv:2107.12930*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

José Cañete. 2019. Compilation of large spanish unannotated corpora.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Maddalen López de Lacalle, Xabier Saralegi, and Iñaki San Vicente. 2020. Building a task-oriented dialog system for languages with no training data: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2796–2802.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Thierry Etchegoyhen and Harritxu Gete. 2020. Handle with care: A case study in comparable corpora exploitation for neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3792–3800. European Language Resources Association.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022a. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawendé François D Assise Bissyande, Jacques Klein, Andrey Boytsov, Anne Goujon, and Clément Lefebvre. 2022b. Luxembert: Simple and practical data augmentation in language model pre-training for luxembourgish. In *Proceedings of the Language Resources and Evaluation Conference, 2022*, pages 5080–5089.

Gati Martin, Medard Edmund Mswahili, Young-Seob Jeong, and Jeong Young-Seob. 2022. Swahbert: Language model of swahili. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, Berlin, Germany. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

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.

David Vilares, Marcos Garcia, and Carlos Gómez-Rodríguez. 2021. Bertinho: Galician bert representations. *arXiv preprint arXiv:2103.13799*.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Nuria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing qubert: A large monolingual corpus and bert model for southern quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel Bowman. 2021. When do you need billions of words of pretraining data? 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 1112–1125.

A Model Pre-training Details

As mentioned in section 2.3, all models were pre-trained following the procedure used for ElhBERTeu (Urbizu et al., 2022). Specifically, We employ a cased sub-word vocabulary of 50K tokens trained with the unigram sub-word segmentation algorithm (Kudo, 2018). We use whole-word masking and train the models for 1M steps with a batch size of 256 and a sequence length of 512 on a single v3-8 TPU for 6 days. Pre-training each model takes about 9.8e+19 FLOPs of computation, and has an estimated 196 kg CO2 emissions, estimated with *Machine-Learning Impact calculator*[7] (Lacoste et al., 2019).

B MT Quality Experiments

B.1 Manual Evaluation of the MT System

The manual evaluation of the MT systems was carried out over a sample containing 100 random sentences extracted from S_beto2eu corpus. The evaluation was done by two bilingual Basque-Spanish speakers. Annotators were presented with the original Spanish sentence and the output produced by the MT system. On the one hand, they were asked to annotate whether the produced sentence was the correct translation of the original, and on the other, whether the generated sentence was linguistically correct. Figure 1 shows the results of the evaluation.

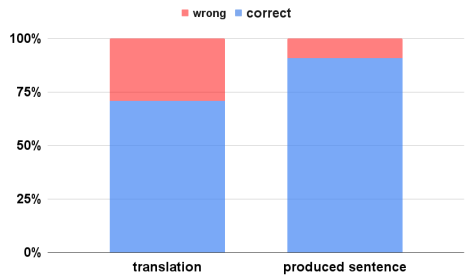

Figure 1: Manual evaluation of the MT system.

B.2 Vocabulary Diversity Experiments

The Basque-Spanish parallel corpus is composed of EITBcc (Etchegoyhen and Gete, 2020) and EhuHac (Tiedemann, 2012) texts. Table 5 presents details of the Basque part of the corpus compared to the Spanish part translated with MT. MT-based translation contains roughly 90K (16%) fewer vocabulary entries.

[7] https://mlco2.github.io/impact#compute

Corpora	Words	Lexicon
Native (EU)	29M	549K
Translated (ES2EU)	28M	462K

Table 5: Lexicon diversity decrement during MT.

C Test Set OOV Analysis

Language Models pre-trained on text from a similar language (dialect, domain or writing style) to the one present in downstream tasks are expected to perform better (Lee et al., 2020; Inoue et al., 2021). An out-of-vocabulary (OOV) analysis was carried out in order to compare the similarity of pre-training data for each language model and the test sets included in the BasqueGLUE Benchmark.

For this OOV analysis, the vocabulary of each test set from BasqueGLUE was compared with the tokenizers of the language models from this work, trained on their respective corpora. The sequential model mentioned in Section 4 is not included in the comparison, as it employs the same tokenizer as ElhBERTeu. The comparison was done at the whole word level for simplicity, excluding words which can be described in several sub-word tokens as overlapping vocabulary.

Table 6 contains the OOV values calculated in this analysis. The table shows that the corpus most similar to the test sets is N_ElhBERTeu, which is the biggest native corpus employed in this work, with an average OOV ratio of 42.82%. ElhBERTeu obtains the lower OOV rates across all the tasks, except for QNLI, where concat50-50 scores lower, and Vaxx and coref, where SN_BERTloc scores lower. Additionally, OOV analysis shows that the corpus S_loc2eu (used to train S_BERTloc) is closer to the test sets than S_beto2eu (used to train S_BERT).

D Results Including StDev

Table 7 collects all the results from Tables 2, 3 and 4 with their respective standard deviations for each task. We can see how some datasets are more stable than others when we finetune a language model on them (models show particularly high standard deviations on Vaxx).

E MLM Results

Table 8 includes evaluations made on Language Modelling. We report the accuracies obtained on MLM, on a news test set not included in the training

of the model.

F Impact of the Amount of Parallel Data

Our approach requires an MT system. While our system is trained on 8.6M parallel sentences, such a parallel corpus is not available for every language. Thus, we analyze if decreasing drastically the amount of parallel data has a notable impact on the resulting MT system quality.

With that aim, we trained an MT system with half of the data. Decreasing the 8.6M parallel sentences to 4M did not have a significant impact on the FLORES-200 benchmark (Team et al., 2022)

Test	NERC	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref	avg
ElhBERTeu	**41.34**	**30.61**	**30.61**	**54.26**	**61.25**	48.84	51.29	**40.94**	26.28	**42.82**
BERT_{125M}	41.77	33.42	33.42	55.71	62.89	49.81	53.07	42.67	26.28	44.34
S_BERT	50.35	35.33	35.33	63.64	67.80	51.81	56.58	48.56	34.66	49.34
SN_BERT	45.05	33.42	33.42	58.96	64.22	49.44	53.45	44.18	29.61	45.75
Sloc_BERT	45.53	34.31	34.31	57.96	63.67	50.88	52.50	44.67	29.04	45.88
SNloc_BERT	41.59	30.87	30.87	54.62	61.47	48.98	**50.56**	40.96	**26.08**	42.89
concat50-50	42.66	30.99	30.99	56.36	62.44	**48.56**	52.07	42.21	27.46	43.75
concat80-20	41.98	32.40	32.40	55.83	62.37	48.79	52.03	42.23	27.12	43.91
paral_{EU}	53.05	41.45	41.45	64.46	67.62	55.80	60.09	53.26	39.18	52.93
paral_{ES2EU}	57.55	47.83	47.83	67.88	71.19	58.64	64.57	58.85	45.02	57.71

Table 6: OOV tokens percentage on the test datasets for the tokenizer vocabulary (as whole words) for each model.

	avg	NERC	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref
ElhBERTeu	73.40±1.20	82.03±0.35	74.13±1.18	**82.19**±1.04	78.48±0.43	69.46±0.43	**76.04**±1.47	59.41±3.21	**72.27**±0.93	66.64±1.78
BERT_{125M}	71.98±1.45	79.51±0.30	75.18±1.35	80.83±1.43	76.94±0.20	70.15±0.94	73.76±2.01	58.32±4.08	70.09±0.70	63.10±2.07
S_BERT	71.40±1.57	79.72±0.40	74.03±1.20	80.94±2.37	73.32±0.54	68.83±1.10	73.93±3.03	58.92±2.32	70.06±1.47	62.83±1.67
SN_BERT	72.38±1.33	**82.33**±0.55	74.12±1.68	81.45±2.08	77.24±0.63	**70.32**±0.59	73.76±1.35	56.31±2.25	71.37±0.82	64.50±2.06
Sloc_BERT	72.21±1.34	79.74±0.71	74.75±1.09	82.26±0.91	75.91±0.16	69.80±0.33	72.66±2.49	61.08±3.32	70.37±1.55	63.27±1.49
SNloc_BERT	72.49±1.53	81.81±0.50	75.14±2.73	80.33±1.85	78.05±0.64	69.95±0.68	72.74±2.64	56.23±2.17	71.63±0.56	66.51±1.98
paral_{EU}	69.19±1.42	73.88±0.70	74.04±0.61	81.82±3.07	72.56±0.82	68.88±1.14	69.87±1.83	53.70±2.11	68.67±0.76	59.28±1.78
paral_{ES2EU}	68.65±1.11	72.42±0.58	72.60±0.50	78.12±0.99	71.09±0.76	68.10±0.67	70.13±1.50	58.74±2.58	68.13±1.21	58.50±1.17
concat50-50	73.12±1.29	81.99±0.42	**75.29**±0.83	79.87±2.04	77.80±0.38	69.23±0.90	72.41±1.22	**62.80**±3.94	71.93±0.70	66.78±1.20
concat80-20	**73.47**±1.28	82.05±0.92	74.21±0.78	81.54±0.79	**78.57**±0.52	68.86±1.02	74.09±2.54	62.17±2.94	71.60±0.72	**68.11**±1.29
sequential	72.75±1.76	81.71±0.75	74.39±1.05	81.55±1.68	77.95±0.37	69.06±1.11	74.09±1.90	57.66±5.33	72.11±1.51	66.24±2.14

Table 7: Results including standard deviations.

	BasqueGLUE	MLM
ElhBERTeu	73.40	61.07
BERT_{125M}	71.98	58.97
S_BERT	71.40	49.67
SN_BERT	72.38	58.64
Sloc_BERT	72.21	53.66
SNloc_BERT	72.49	60.42
paral_{EU}	69.19	43.75
paral_{ES2EU}	68.65	41.98
concat50-50	73.12	59.84
concat80-20	73.47	61.05
sequential	72.75	58.61

Table 8: Results for MLM accuracies.

parallel data	BLEU	charF++
8.6M sentences	13.2	47.4
4M sentences	13.0	47.2

Table 9: Results for the MT data ablation study. MT evaluation at FLORES-200, we report BLEU and charF++.

obtaining 13.0 BLEU (-0.2) and 47.2 chrF++ (-0.2) (see Table 9). Hence with a similar MT quality, we expect a comparable performance of LM models.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations, after 5 Conclusions

☒ A2. Did you discuss any potential risks of your work?
No, we are not aware of any potential risks of our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and 1 Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2 Methodology, 3 "LM Trained Solely on Synthetic Data", and "4 Combining Native and Synthetic Data"

☑ B1. Did you cite the creators of artifacts you used?
2 Methodology

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2 Methodology and "4 Combining Native and Synthetic Data" and appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2 Methodology and appendix

C ☑ Did you run computational experiments?

3 "LM Trained Solely on Synthetic Data", "4 Combining Native and Synthetic Data" and appendix

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
2 Methodology and appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2 Methodology and appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
2 Methodology and appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
2 Methodology

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.