TEMA: Token Embeddings Mapping for Enriching Low-Resource Language Models

Rodolfo Zevallos
Universitat Pompeu Fabra
Barcelona, Spain
rodolfojoel.zevallos@upf.edu

Núria Bel
Universitat Pompeu Fabra
Barcelona, Spain
nuria.bel@upf.edu

Mireia Farrús
Universitat de Barcelona
Barcelona, Spain
mfarrus@ub.edu

Abstract

The objective of the research we present is to remedy the problem of the low quality of language models for low-resource languages. We introduce an algorithm, the Token Embedding Mapping Algorithm (TEMA), that maps the token embeddings of a richly pre-trained model L1 to a poorly trained model L2, thus creating a richer L2' model. Our experiments show that the L2' model reduces perplexity with respect to the original monolingual model L2, and that for downstream tasks, including SuperGLUE, the results are state-of-the-art or better for the most semantic tasks. The models obtained with TEMA are also competitive or better than multilingual or extended models proposed as solutions for mitigating the low-resource language problems.

1 Introduction

At the forefront of artificial intelligence, large language models have achieved extraordinary levels of performance in most natural language processing tasks. These advancements, driven by deep learning techniques, highlight the ability of these models to process complex linguistic contexts. However, their effectiveness seems to be intrinsically linked to the enormous amount of data required for pre-training, thus raising concern about how low-resource languages could benefit from them. Low-resource languages, that is, languages that do not have a massive amount of text, risk being almost excluded from the possibility of having good NLP applications. For coping with the problems of low-resource languages in large language models, there have been different proposals, mainly: building multilingual models and applying transfer learning techniques and the so-called data augmentation methods.

Augmentation methods for language modelling are those that automatize the production of new texts by different means: template-based scripts (Wei and Zou, 2019), machine translation (Sennrich et al., 2016), and delexicalized seq2seq models (Hou et al., 2018; Zevallos et al., 2022a), for instance. However, it is costly to guarantee the semantic correctness of the produced texts. Multilingual models are models trained with texts in different languages (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020a). In addition to accepting different languages, they can then be fine-tuned on a downstream task using labelled data in only one language and it is expected to generalize in order to handle samples of the other languages it has been pre-trained with. However, low-resource languages are still under-represented in the vocabulary impairing proper modelling.

In this context, we propose to remedy the low quality of language models due to poor pre-training data by translating (i.e. mapping) token embeddings of a richly trained model L1, to the position of the translation equivalent token embedding in a poorly pre-trained model L2, thus creating a richer L2' model. Our Token Embedding Mapping Algorithm (TEMA) works as if the learned lexical parameters of L1 were transferred to L2, so that even tokens of L2 for which only one example is available get a rich representation. We assessed the efficacy and the impact of TEMA with different experiments with Quechua, which is a low-resource language and, in addition, it is a polysynthetic language with more than 100 inflectional suffixes. Languages with inflectional or synthetic morphology have a larger number of different tokens than other languages such as Chinese or English and are harder to model (Mielke et al., 2019; Zevallos and Bel, 2023). In addition, and in order to be able to evaluate the impact of using TEMA with standard benchmarks, we provide the results of working with English, German and French for which evaluation datasets are available and for which we have reproduced poor pre-training conditions. Finally, we compare the results of our approach with other

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11423–11435
November 12-16, 2024 ©2024 Association for Computational Linguistics

cross-lingual transfer methods and with standard multilingual models, which are standard proposed solutions for having models for low-resource languages.

2 Related Work

Multilingual models and data augmentation methods have been proposed as solutions for transformer-based language models to satisfactorily process low-resource languages.

For instance, the multilingual mBERT (Devlin et al., 2019) was trained with data belonging to 104 different languages and no special supervision like parallel corpora or bilingual dictionaries. The resulting multilingual representations were said to be shared tokens, mostly subwords, for many different languages. These results encouraged the research on transfer learning techniques (Alyafeai et al., 2020) under the assumption that pre-training a very large model with many languages is to get representations with general-purpose "knowledge" that would improve the performance on downstream tasks for many different languages. However, Conneau et al. (2020a) already show that performance degraded across all languages and that low-resource languages are under-represented in the vocabulary, which mostly contains the tokens of the largest languages. The common tokens that multilingual models use as crosslingual token embeddings to transfer information from one language to another are at the end subwords made of small groups of characters for which the model cannot learn good representations.

Artetxe et al. (2020) demonstrated that cross-lingual learning transfer to get a shared vocabulary could be done without joint training. Their method transferred a monolingual model L1 to a new language L2 by only substituting token embeddings of L1 with the new token embeddings learnt by further training the monolingual model with data of the new language L2, while freezing the transformer. To evaluate the impact of transferring the model from one language to another, among others, these authors used the WiC dataset (Word in Context, (Pilehvar and Camacho-Collados, 2019)) delivering competitive results. Dobler and de Melo (2023) and Minixhofer et al. (2021) propose to use token embeddings from a model trained with abundant resources as initialization embeddings for a low-resource model. Our method keeps the idea of working only at the lexical level, but instead of further training with L2 data, transferring the lexical knowledge, i.e. projecting the token embeddings, from a richly trained L1 to a poorly pre-trained monolingual model of L2.

In the method we are presenting, the transfer is done by a mapping function inspired by other previously proposed methods. Mikolov et al. (2013) and Lample et al. (2018) worked under the hypothesis that words from different languages get similar static embedding representations. Therefore, to align the monolingual spaces should be possible with minimal supervision, for instance, common tokens, to create bilingual dictionaries automatically with a mapping function (Artetxe et al., 2017; Zhang et al., 2017; Conneau et al., 2017).

The idea of automatically producing texts, or data augmentation, was first used to improve system robustness. Jia and Liang (2017) or Ribeiro et al. (2018) proposed to evaluate the understanding capacities of systems by creating synthetic data, following the idea from computational vision, where, in order to test a learner's oversensitivity, a bit of noise is added to the test sample to evaluate whether the system changes the decision taken for the original test input. On the same basis, there were different proposals to automatically generate semantically equivalent sentences using, for instance, paraphrase generation techniques. Mallinson et al. (2017) method was based on Neural Machine Translation back-translation methods. Once paraphrases were found, a number of heuristic pattern-matching rules were induced, and later filtered to produce a set of rules to automatically produce new sentences from any given reference dataset. Other methods for creating new sentences are based on: word-level modifications of original sentences based on synonym replacement (Wei and Zou, 2019), using LSTM language models (Kobayashi, 2018), Multilingual Language Modelling (Conneau et al., 2020b) and auto-regressive pre-trained Language Models (Kumar et al., 2020). However, these methods suffer from generating semantically unrealistic sentences that have to be manually revised in most cases. Proposals of methods to preserve semantics are Hou et al. (2018), Zhou et al. (2019), or Zevallos et al. (2022a) that used a delexicalized seq2seq model for generating some similar sentences each with different words of the same semantic field as found in resources like Wordnet (Fellbaum, 1998; Melgarejo et al., 2022). Eventually, this method allows for increasing the vocabulary of the model, a side effect that has proved to be beneficial in all the scenarios.

Enhancing vocabulary is one of the benefits highlighted in eB-BERT (Wang et al., 2020). eB-BERT consists of a bilingual model, B-BERT, which has an extended vocabulary and is fine-tuned with data from the target low-resource language. The vocabulary extension involves adding the 10,000 most frequent tokens from the low-resource language to the original B-BERT vocabulary. The fine-tuning process uses only the masked language modeling objective with the L2 data. This method was proposed to avoid training a multilingual model from scratch with a new language, addressing the issue that in a multilingual model, there is often limited space in the vocabulary for tokens from languages with smaller training datasets. Hangya et al. (2022) proposed to fine-tune a cross-lingual model with mined translation word pairs extracted from two monolingual language models, and confirmed that extending the vocabulary was one of the keys for improving the coverage of low-resource languages.

Also close to our approach is the line of research inspired by the word replacement strategies by Zhou et al. (2021), which proposed a virtual data augmentation method for fine-tuning pre-trained monolingual language models. In embedding augmentation, the original token embeddings are replaced by ones which are derived by a probabilistic mixture of the embeddings of the vocabulary terms that are predicted by the masked language model conditioned on the input. Our approach also substitutes specific tokens created in a monolingual language model but with the embeddings created in another monolingual model that however was trained with a larger amount of data.

3 Token Embedding Mapping

We propose to remedy the low quality of a language model poorly pre-trained L2 by mapping its token embeddings to the position of its translation equivalents in a richly pre-trained model L1. The Token Embedding Mapping Algorithm (TEMA) we propose works as if the learned lexical parameters of L1 were transferred to L2, so that even tokens of L2 for which only one example is available get a rich representation.

Figure 1 illustrates how mapping token embeddings is done by translating token embeddings of a poorly trained model L2 (u_n) to the position of the translation equivalent token embedding in a richly pre-trained model L1 (w_m), thus creating a richer L2’ model. To do this, we employed the affine transformation method (Yang et al., 2021) using as references the tokens of the number ’1’ (v_x and v_y) which exists in the two models, a method proposed in research on inducing bilingual word embeddings (Artetxe et al., 2017).

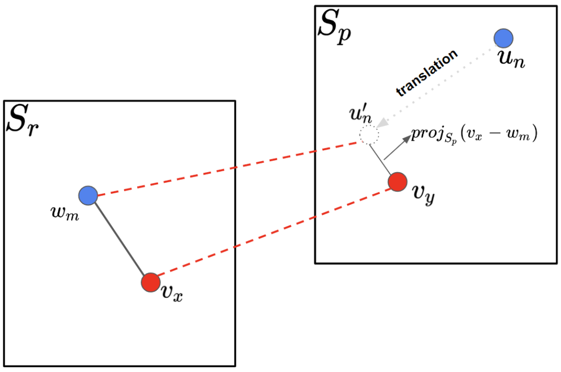

Figure 1: Being S_r the vector space of the richly pre-trained model L1, and S_p the one of the poorly pre-trained model L2, TEMA translates token embeddings like u_n to u'_n, that is the projection of u_n equivalent token in L1, w_m, taking as references the token embeddings v_x and v_y, that correspond to the token ’1’ in each models

In equation 1, u'_n represents the enriched version of u_n, and $\text{proj}_{S_p}(v_x - w_m)$ is the projection of the vector $v_x - w_m$ in the L2 vector space S_p.The projection $\text{proj}_{S_p}(v_x - w_m)$ onto S_p using v_y as a reference is obtained with equation 2:

$$u'_n = u_n + \text{proj}_{S_p}(v_x - w_m) \tag{1}$$

$$\text{proj}_{S_p}(v_x - w_m) = \frac{v_y(v_x - w_m)}{||v_x - w_m||^2}(v_x - w_m) \tag{2}$$

3.1 TEMA

In this section, we describe TEMA (Token Embedding Mapping Algorithm). The algorithm takes the reference token embeddings v_x and v_y as input from L1 and L2, respectively, as well as two lists of tuples D_x and D_y. These tuples contain the index of two translation-equivalent tokens (i_m and j_n, respectively), where each index is associated with a word in the corresponding language, along with their respective embeddings (w_m and u_n).

The algorithm utilizes the function T that identifies token pairs between L1 and L2 translation equivalents as found in a bilingual dictionary. TEMA begins by using T, and in case it returns

$u_n = \emptyset$, we add the target token (tgt_token) using the index j_n from D_y to the list V, which contains all tokens missing in the model L2. We repeat this process until we traverse all of D_x. If $V \neq \emptyset$ then for each token in V, we add a sentence (which can be an example phrase from the dictionary or other sources) to the file named tgt_sent. After adding all the phrases, we fine-tune the model L2 with the tgt_sent file, resulting in the model L2'.

Finally, we iterate again through each i_m and w_m in D_x. In each iteration, we use T, and if T returns $u_n \neq \emptyset$, we calculate the projection using equations 1 and 2. Once we obtain u'_n, we update D_y with it. When the iteration finishes, we update the model L2' (or L2 if $V = \emptyset$) with D_y.

3.1.1 Model L2 update

To update the model L2 with the modified token embeddings from TEMA, we perform fine-tuning of the model by adding an embedding layer. In this way, the parameters of the pre-training model are retained while incorporating the modified embeddings. The process is similar to fine-tuning for a downstream task, so we have decided to use the same hyperparameters as those employed in a Part-of-Speech tagging task. Importantly, by introducing an embedding layer during fine-tuning, we can effectively update the L2 model to utilize the token embeddings that have been modified by TEMA. This approach maintains the pre-trained weights of the model's core layers, enabling efficient adaptation to the new embedding space.

This step is included in the code that we will provide regarding TEMA and our experiments.

4 Experimental Setup

Our main target was the use of TEMA for modelling Quechua, which is a low-resource language. Additionally, as already mentioned, in order to accurately evaluate its impact with standard benchmarks and compare with other methods, we designed different experiments involving another three L2-languages, for which we reproduced low-resource conditions. For the methods that involve two languages, we paired them as follows: L1-Spanish and L2-English (es-en), L1-English and L2-German (en-de), L1-English and L2-French (en-fr), and L1-Spanish for the L2-Quechua (es-qu). In the subsections 4.1 and 4.2, we describe the resources used: bilingual dictionaries and training corpora, in subsection 4.4 we describe the different language models that have been trained for the experiments and in subsection 4.3 the used tokenizers: BPE and DeepSpine.

As for the evaluation, we provide results of intrinsic evaluation in terms of pseudo-perplexity (Salazar et al., 2019), and of different downstream tasks as extrinsic evaluation.

4.1 Bilingual Dictionaries

For our experiments involving the translation equivalents for Quechua, we used the Quechua-Spanish Bilingual Dictionary (Calvo-Pérez, 2022). This dictionary is one of the most comprehensive and up-to-date, available for both varieties of Quechua (Chanka and Collao) and it comprises 74,395 Spanish entries and 51,233 Quechua entries.

The data for the Spanish-English, English-German, and English-French pairs comes from the respective bilingual digital dictionaries in "wiktionary.org". "Wiktextract" tool (Ylonen, 2022) facilitated the extraction of all entries for the language pairs used in our experiments in a rather easy and quick manner.

4.2 Corpora

For our experiments, we trained different models (BERT, RoBERTa, B-BERT) with datasets of two different sizes: 1B for the L1 Spanish and English richly trained models and 10M for the L2 monolingual poorly trained models. As training data for the Spanish 1B model, we created a 1 billion-word corpus from the MarIA model training corpus (Gutiérrez-Fandiño et al., 2021), which comprises 133B tokens from texts crawled from .es domains by the National Library of Spain. Similarly, to train the 1B English models, we utilized a 1 billion-word portion of the BERT training corpus from Devlin et al. (2019), which includes the English Wikipedia (2.5B tokens) and BookCorpus (800M tokens) (Zhu et al., 2015).

The 10M models trained for English, German, French, and Quechua used the following resources. The L2 English models used the same corpus source as the L1 models of the same language but only with 10M words. For both German and French, we used the OSCAR corpora extracted from Common Crawl (Ortiz Suárez et al., 2019), which consist of 21B and 32.7B words, respectively. Finally, for Quechua, we used the latest version of the Monolingual-Quechua-iic corpus[1], which com-

[1] https://huggingface.co/datasets/Llamacha/monolingual-quechua-iic

Model	P	V-B	V-D
BERT	110M	30k	30k
RoBERTa	125M	50k	36k
RoBERTa + TEMA	125M	50k	36k
eB-BERT	110M	120k	63k
eB-BERT + lus	110M	120k	63k

Table 1: Hyperparameters for all types of language models used in our experiments are consistent across all languages. Each type of language model shares the same configuration: featuring 12 layers and attention heads, and a feed-forward network dimension of 3072. Additionally, P denotes the number of parameters; V-B represents the number of BPE vocabulary items; V-D represents the number of DeepSpin vocabulary items.

prises 10M words and was employed by Zevallos et al. (2022b) to develop Quechua language models. This Quechua corpus draws from diverse sources, including Wikipedia (approximately 1M tokens), other online resources, educational materials, and legal documents.

4.3 Tokenizers

TEMA is based on translation-equivalent words as found in a bilingual dictionary, and therefore, tokenization becomes a critical point. Because Byte Pair Encoding (BPE) (Sennrich et al., 2015), the most used tokenizer, looks for frequent character sequences, it creates many subwords that cannot be found in a standard bilingual dictionary, and this problem gets worse when dealing with a limited number of texts. Thus, for our experiments, we compared the results of using BPE and of using a supervised tokenizer, DeepSpin (Peters and Martins, 2022), that produces linguistically motivated subwords and delivers stem-like tokens that are more likely to match entries in a bilingual dictionary. For Quechua, we used DeepSpin modules as described in (Zevallos and Bel, 2023).

4.4 Baseline Models

We trained several language models from scratch for both L1 and L2 languages (with data as described in Section 4.2) using two types of language models: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and two other methods aiming at language modeling for low-resource languages: the extended B-BERT (eB-BERT) by Wang et al. (2020) and the "linear_unsup" (lus) model by Hangya et al. (2022). Finally, we also compared TEMA's results on downstream tasks with off-the-shelf multilingual base models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a) fine-tuned for Quechua with the data just described.

4.4.1 Monolingual Base Models

We pre-trained all our L1 and L2 language models from scratch using BPE and DeepSpin tokenizers and following the recipes of the bert-base-cased[2] and roberta-base[3] available on Hugging Face.

In total, we pretrained 16 models, comprising four BERT-BPE, four BERT-DeepSpin, four RoBERTa-BPE, and four RoBERTa-DeepSpin. Details of the size and hyperparameters used in the BERT and RoBERTa model types are given in Table 1 and Table 5 (in the appendix).

4.4.2 RoBERTa + TEMA

TEMA was used with RoBERTa models for L1 and L2, and the resulting models are named RoBERTa + TEMA for each language pair utilizing both BPE and DeepSpin.

4.4.3 Bilingual fine-tuned models

In order to compare TEMA results with the ones obtained by other methods that are proposed to improve the performance of language models for low-resource languages, we used eB-BERT (Wang et al., 2020) and eB-BERT + linear_unsup (Hangya et al., 2022).

For eB-BERT training, we followed Wang et al. (2020) only adding more vocabulary when retraining BERT with data of a low-resource language. In case of eB-BERT + linear_unsup, we used it with our languages following the steps described in Hangya et al. (2022). In total, we trained 16 models: eight eB-BERT models (four BPE and four DeepSpin) and eight eB-BERT + linear_unsup models (four BPE and four DeepSpin) thus covering the following language pairs: es-en, en-de, en-fr and es-qu. Details of the size and hyperparameters used in the bilingual fine-tuned model types are given in Table 1 and Table 6 (in the appendix).

4.4.4 Multilingual models

In addition to comparing RoBERTa + TEMA with the monolingual and bilingual models described above, we also compared with two off-the-shelf

[2] https://huggingface.co/google-bert/bert-base-cased

[3] https://huggingface.co/nyu-mll/roberta-base-1B-3

Tokenizer	BPE				DeepSpin			
Monolingual base models / L2	en	de	fr	qu	en	de	fr	qu
BERT 10M	51.2	83.4	98.6	391.2	17.8	36.1	49.5	141.3
RoBERTa 10M	10.8	13.8	17.1	245.1	8.1	10.3	12.4	85.1
TEMA / L1 → L2	es-en	en-de	en-fr	es-qu	es-en	en-de	en-fr	es-qu
RoBERTa 1B & RoBERTa 10M	**4.8**	**9.7**	**10.6**	**58.3**	**4.5**	**7.9**	**8.5**	**21.1**
Bilingual fine-tuned models / L1 + L2	es-en	en-de	en-fr	es-qu	es-en	en-de	en-fr	es-qu
eB-BERT 1B + 10M	9.6	13.6	19.4	155.9	9.0	10.5	14.8	69.3
eB-BERT 1B + 10M + linear_unsup	8.3	10.7	15.1	96.7	8.2	9.9	11.7	58.7

Table 2: Pseudo-perplexity results of monolingual, bilingual and TEMA models on different languages and tokenizers.

multilingual models: mBERT[4] (Devlin et al., 2019) and XLM-RoBERTa[5] (Conneau et al., 2020a). In the case of Quechua, we had to fine-tuned mBERT and XLM-RoBERTa with Quechua-L2 because these two models were not pre-trained with a Quechua corpus.

4.5 Evaluation: pseudo-perplexity and downstream tasks

We first evaluated the different models in terms of perplexity. Additionally, we used two widely used datasets in language model evaluation: Xtreme (Hu et al., 2020) and SuperGLUE (Wang et al., 2019).

We evaluated all models through fine-tuning across the following Xtreme tasks: (1) Classification (XNLI and PAWS-X), (2) Structure prediction (POS and NER), and (3) Question answering (XQuAD). For the experiments, we adhered to the recipes[6] provided by Xtreme for fine-tuning each task for every language (see Table 7 in the appendix).

As for SuperGLUE, due to the lack of multilingual data, the fine-tuning was only with the English models and we selected five representative tasks from SuperGLUE: BoolQ, CB, Copa, RTE, and WiC, following the methodology outlined by Wang et al. (2019). We fine-tune using an epoch of 5 and a batch_size of 16 in all cases.

5 Results

5.1 Pseudo-perplexity

We present the pseudo-perplexity results in Table 2. RoBERTa + TEMA are the models that achieved the best performance across all language when using BPE and DeepSpin. In the particular case of Quechua, the reduction is significant, decreasing from 391.2 to 21.1, when using RoBERTa + TEMA and DeepSpin.

Models	BoolQ	CB	Copa	RTE	WiC
Monolingual base models / L2					
BERT	0.65	0.65	0.61	0.55	0.52
RoBERTa	0.68	0.74	0.78	0.69	0.57
TEMA / L1 → L2					
R+TEMA	0.77	0.78	0.80	0.74	0.63
Bilingual fine-tuned models / L1 + L2					
eB-BERT	0.72	0.73	0.66	0.63	0.57
eB-BERT+lus	0.74	0.75	0.67	0.66	0.60

Table 3: Results (accuracy) on the SuperGLUE dataset for English among monolingual and bilingual and RoBERTa + TEMA models. R+TEMA = RoBERTa+TEMA.

5.2 Downstream Tasks

Table 8 (in the appendix) shows that the RoBERTa + TEMA models obtained improvements over the monolingual and bilingual models in all languages for all the Xtreme tasks. The effectiveness of TEMA algorithm is best appreciated in the results for Quechua in PoS and NER, managing to outperform all models (including mBERT and XLM-RoBERTa).

Table 3 illustrates that RoBERTa + TEMA significantly[7] outperforms both monolingual and bilingual models; Specifically, in the WiC task, which involves semantics and the distribution of token embeddings in a vector space, RoBERTa + TEMA increases the average accuracy by approximately 0.11 compared to baseline monolingual models and by about 0.03 compared to bilingual models.

[4] https://github.com/google-research/bert/blob/master/multilingual.md

[5] https://huggingface.co/FacebookAI/xlm-roberta-base

[6] https://github.com/google-research/xtreme

[7] Improvements are statistically significant at $p < 0.05$ as assessed with a sign test

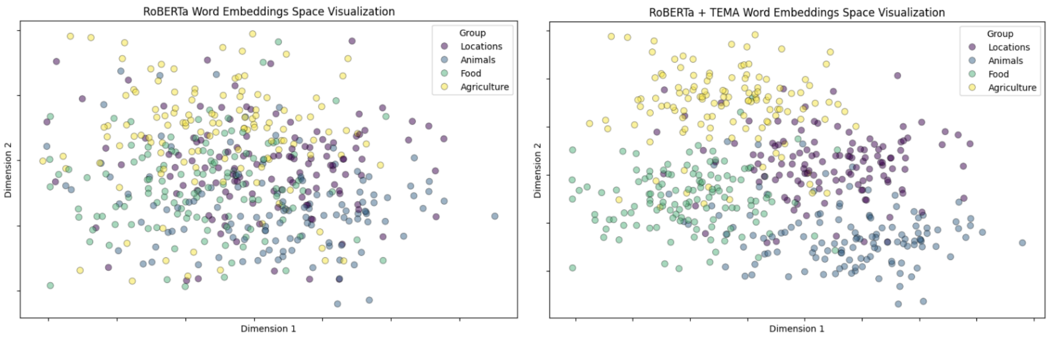

Figure 2: The left graph plots the vector space of the Quechua RoBERTa model. The right one plots the Quechua RoBERTa + TEMA one showing that, after translating the embeddings, the extra information results in a more semantic vector space. The coloured points belong to different semantic groups: locations, animals, food and agriculture, and they appear to be better represented after applying TEMA.

6 Discussion

Results in all the experiments show that the affine transformations on the vector space of the model are not causing trouble. On the contrary, the experiments we have carried out validate that TEMA indeed mitigates the issues due to the low quality of the representations of low-resource languages.

First, low-resource languages have problems finding enough training data in terms of the total amount of tokens, that is, the vocabulary, but also in the number of samples of each token. Token frequency has proved to be critical in transformer-based language modelling. For instance, Kassner et al. (2020) demonstrated that when a token appears fewer than 15 times in the training data, a BERT model will disregard it, while a token that appears 100 times or more will be predicted more accurately. Wei et al. (2021) found evidence too that BERT models cannot correctly predict tokens that occur less than 10 times in the corpus. Therefore, for low-resource languages, many of the tokens occurring in the corpus do not achieve the frequency required for the model to represent them correctly.

In the results shown in Table 2, we can see an improvement in the perplexity score when using TEMA for all the languages trained with only 10M, and crucially with few sentences per vocabulary token. For instance, for the Quechua RoBERTa + TEMA model, 375 new words were added with 375 single samples, while for English 2.573 new words were added. Note that eB-BERT could be said to be, in spirit, similar to our TEMA. Both add vocabulary of the low-resource language. eB-BERT adds 10k new tokens. The difference is that TEMA also improves its representation when projecting the word embedding of the translation equivalent token that has been obtained with many more examples.

Perplexity reduction is notable for all languages. Note that, as expected, a tokenization that delivers subwords that are more similar to word stems improves TEMA results for all the languages. These modelling improvements can be further appreciated with the fill-mask tests for the es-qu RoBERTa + TEMA model compared to the Quechua RoBERTa 10M model. Table 4 shows that the word probabilities produced by RoBERTa + TEMA makes more sense in each of its sentences than the ones delivered by RoBERTa.

Second, the semantic quality of the token representations is improved using TEMA. Figure 2 shows a sample of the vector space of Quechua models with tokens related to different semantic groups: locations, animals, food and agriculture-related words. The improvement after TEMA is applied is that the distance between related words is reduced, showing a better-structured space that should better support semantic tasks.

Indeed, the results of the more semantic SuperGLUE downstream tasks validated this hypothesis. Results in Table 3 show accuracy improvements in SuperGLUE tasks when using TEMA. In particular, WiC task is very sensitive to lexical and contextual information, as it is deciding whether a token has been used with the same sense in two different sentences. The improvements achieved for the WiC task are evidence of TEMA creating a better vector space.

The improvement in the semantic quality of the embeddings is also evident from the examples from

Sentence	RoBERTa		RoBERTa + TEMA	
	Word	Score	Word	Score
<mask> tutamanta laq'akun. (<mask> barks at night.)	Wawaqa (the boy)	0.4471	Allquqa (the dog)	0.8731
	Runaq (the man)	0.2511	Atoqqa (the wolf)	0.1103
	Allquqa (the dog)	0.2281	Misiqa (the cat)	0.0165
Carlosqa tiendamanta t'antata <mask>. (Carlos <mask> bread from the store.)	Garcia	0.3219	rantisq (bought)	0.1682
	wayk'un (bakes)	0.1814	rantin (buy)	0.1104
	qhawan (looks)	0.0836	ruwan (makes)	0.0918
Chay waynaqa Lima llaqtamanmi <mask> estudiananpaq. (The boy <mask> to Lima to study.)	ripun (goes)	0.1528	ripun (goes)	0.3581
	purin (travels)	0.1011	hamun (comes)	0.1810
	wicharin (goes up)	0.08131	rirqa (went)	0.1555

Table 4: Exploring Word Probability (Fill-Mask) in RoBERTa 10M (Quechua L2) vs RoBERTa + TEMA (L1 and L2).

the fill-mask test shown in Table 4. In all cases, applying TEMA improves the results by assessing as more probable the more semantically related words given the sentence context.

Finally, for low-resource languages, the use of massive multilingual models has been proposed to be the solution. However, the criticism of general underperformance is confirmed by the results in Table 8. Results for the Xtreme tasks are in line with the resources of each particular language used for pre-training. English achieved best results with XLM-R after using more than 55B tokens training corpus. German and French, with training corpora around 10B each, also showed better results than other models trained only with 10M, but not as good as one could expect for tasks such as PoS tagging reaching 0.88 F1. For a new language which is not among the 100 languages in the multilingual training data, TEMA is by far the best choice, even better that a multilingual model. Note that for Quechua, XLM-R achieved only 0.72 F1 for PoS, while RoBERTa + TEMA achieved 0.84 F1, and similar better results for NER are also shown in Table 8.

7 Conclusions

We have introduced TEMA: Token Embedding Mapping Algorithm, a method for mapping the token embeddings of a richly pre-trained language model to a poorly pre-trained model. In this paper, we have provided experimental evidence that the language models resulting from using TEMA for four different languages showed reduced perplexity and very competitive performance in Xtreme and SuperGlue tasks compared to other systems based on transfer techniques and lexicon expansion methods. Another contribution of our experiments is the confirmation that the amount of lexical items of a language in the vocabulary is directly affecting the results. However, our experiments demonstrate that TEMA crucially transfers adequately the lexical parameters learned with larger data, getting better results by overcoming the problem of the amount of data available for each token in a small corpus, or after just enlarging the lexicon. Therefore, TEMA contributes to the research on NLP methods for low-resource languages by proposing a method that only requires a bilingual dictionary and a rather small monolingual corpus of 10M tokens to deliver state-of-the art results or even better for basic NLP tasks such as PoS tagging and NER. Resources and code are available at https://github.com/IULATERM-TRL-UPF/TEMA

8 Limitations

Although our results demonstrate improvements in both pseudo-perplexity and downstream tasks, the most notable limitations of our work revolve around the availability of bilingual dictionaries between a language with rich resources and one without. Despite the mentioned difficulty, acquiring a bilingual dictionary is easier than obtaining a parallel corpus to work with.

On the other hand, given the limitations of BPE for handling morphologically complex languages, such as Quechua, pre-training large models (L1 in our case) with linguistically motivated tokenizers is needed to fully leverage the use of TEMA. Furthermore, this requirement comes with computational limitations, as training an L1-model from scratch with all its resources is practically impossible for many researchers.

Finally, another significant limitation to mention is the amount of vocabulary in the training corpus, as the more vocabulary the corpus has, the better the use of TEMA. However, this can be mitigated with a bilingual dictionary that provides a large number of entries for the low-resource language.

9 Ethical considerations

We could not identify any specific ethical issue in this paper or potential danger. However, it is crucial to emphasize the significant importance of working with low-resource languages, which might entail ethical complexities that need to be carefully addressed in natural language processing (NLP). These languages, often marginalized or underrepresented, pose unique challenges in terms of data availability, tools, and linguistic resources. Tackling these complexities ethically and equitably is essential to ensure the proper inclusion and representation of these languages. To delve deeper into the complexities of Quechua, we recommend further reading (Camacho Caballero and Zevallos Salazar, 2020).

Acknowledgements

This research was partially funded by LUTEST Project PID2019-104512GB-I00, Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación (Spain). The first author has been supported by a FI grant of the Catalan Funding Agency for Research and Universities (AGAUR)

References

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Julio Calvo-Pérez. 2022. *Nuevo diccionario Español-Quechua, Quechua-español*. Universitat de València.

Luis Camacho Caballero and Rodolfo Zevallos Salazar. 2020. Lingüística computacional para la revitalización y el poliglotismo. *Letras (Lima)*, 91(134):184–198.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual language model pretraining*. Curran Associates Inc., Red Hook, NY, USA.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv'e J'egou. 2017. Word translation without parallel data. *ArXiv*, abs/1710.04087.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? *arXiv preprint arXiv:2006.10413*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Nelsi Melgarejo, Rodolfo Zevallos, Héctor Gómez, and John E Ortega. 2022. Wordnet-qu: Development of a lexical database for quechua varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2021. Wechsel: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. *arXiv preprint arXiv:2112.06598*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low-resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Ben Peters and Andre F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–138, Seattle, Washington. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Jinfa Yang, Yongjie Shi, Xin Tong, Robin Wang, Taiyan Chen, and Xianghua Ying. 2021. Improving knowledge graph embedding using affine transformations of entities corresponding to each relation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 508–517.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Rodolfo Zevallos and Nuria Bel. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics.

Rodolfo Zevallos, Núria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022a. Data Augmentation for Low-Resource Quechua ASR Improvement. In *Proc. Interspeech 2022*, pages 3518–3522.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022b. 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, Hybrid. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. 2019. Unsupervised context rewriting for open domain conversation. 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 1834–1844, Hong Kong, China. Association for Computational Linguistics.

Kun Zhou, Wayne Xin Zhao, Sirui Wang, Fuzheng Zhang, Wei Wu, and Ji-Rong Wen. 2021. Virtual data augmentation: A robust and general framework for fine-tuning pre-trained models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3875–3887, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Appendix

A.1 Model and training procedure: details

A.1.1 Pre-trained Language Models

We pre-trained all language models (1B and 10M) on 5 server equipped with an Intel Xeon E5-2650 v4 CPU (12 cores, 2.2GHz 30MB Cache 2400MHz 105W) and a Gigabyte Geforce GTX 1080 Ti TURBO 11.72GB GPU. We pre-trained the L2 (10M) models for 10k steps. The training time was over 4 days for models. On the other hand, the L1 (1B) models were pre-trained using 100k steps and took 14 days for models. The entire experiment took approximately 32 days. Table 5 and Table 6 are described all hyperparameters.

	10M			**1B**		
	B	**R**	**R+T**	**B**	**R**	**R+T**
L	12	6	6	12	12	12
H	512	512	512	768	768	768
AH	12	8	8	12	12	12
Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Warmup Steps	10k	6k	6k	10k	6k	6k
Learning Rates	5e-5	5e-4	5e-4	5e-5	5e-4	5e-4
Batch	256	512	512	256	1024	1024
Weight Decay	0.1	0.1	0.1	0.1	0.1	0.1
Max Steps	10k	10k	10k	10k	31k	31k
Adam e	1e-4	1e-6	1e-6	1e-4	1e-6	1e-6
Adam B_1	0.9	0.98	0.98	0.9	0.98	0.98
Adam B_2	0.999	0.98	0.98	0.999	0.98	0.98

Table 5: Hyperparameters for pretraining monolingual models.

Hyper-parameter	eB-BERT	eB-BERT_lus
Number of Layer	12	12
Hidden size	512	512
Attention Heads	12	12
Dropout	0.1	0.1
Warmup steps	10k	10k
Learning Rates	2e-5	2e-5
Batch Size	32	32
Weight Decay	0.1	0.1
Max Steps	500k	500k
Adam e	1e-4	1e-4
Adam B_1	0.9	0.9
Adam B_2	0.999	0.999

Table 6: Hyperparameters for pretraining bilingual models.

A.1.2 Downstream Task

Similar to pre-trained language models, we used 5 server equipped with an Intel Xeon E5-2650 v4 CPU (12 cores, 2.2GHz, 30MB Cache, 2400MHz, 105W) and a Gigabyte GeForce GTX 1080 Ti TURBO 11.72GB GPU to train all these models. The training time was approximately 5 hours per task (Xtreme and SuperGLUE). In total, we trained models over a period of approximately 8 days.

Task	B	E	LR	G_ACC
XNLI	16	5	2e-5	4
PAWS-X	16	5	2e-5	4
POS	8	20	2e-5	4
NER	8	20	2e-5	4
XQuAD	16	3	3e-5	8

Table 7: Hyperparameters used in the tasks of Xtreme for all languages. Batch-size=B, Epoch=B, Learning-rate=LR, Gradient_accumulation_step=G_ACC.

A.2 Experiment results

Languages (L1-L2)	Pair Sentence (Hu et al., 2020)		Structured prediction (Hu et al., 2020) (Zevallos et al., 2022b)		QA (Hu et al., 2020)
	XNLI	PAWS-X	POS	NER	XQuAD
Metrics	Acc.	Acc.	F1	F1	F1
Monolingual base models / L2 (10M)					
			BERT		
en	0.72	0.78	0.88	0.74	0.69
de	0.64	0.67	0.80	0.65	0.61
fr	0.61	0.68	0.75	0.67	0.66
qu	-	-	0.66	0.62	-
			RoBERTa		
en	0.73	0.80	0.89	0.76	0.70
de	0.65	0.67	0.81	0.65	0.62
fr	0.63	0.69	0.76	0.68	0.67
qu	-	-	0.66	0.63	-
TEMA / L2 → L1					
			RoBERTa + TEMA		
es-en	0.78	0.84	0.92	0.80	0.76
en-de	0.70	0.73	0.83	0.71	0.68
en-fr	0.67	0.73	0.80	0.70	0.71
es-qu	-	-	**0.84**	**0.75**	-
Bilingual fine-tuned models / L1 + L2					
			eB-BERT		
es-en	0.74	0.80	0.89	0.75	0.70
en-de	0.65	0.68	0.81	0.66	0.62
en-fr	0.63	0.68	0.76	0.67	0.68
es-qu	-	-	0.71	0.65	-
			eB-BERT + lus		
es-en	0.77	0.82	0.92	0.79	0.73
en-de	0.68	0.70	0.83	0.69	0.65
en-fr	0.65	0.71	0.79	0.69	0.70
es-qu	-	-	0.76	0.70	-
Multilingual full models					
			mBERT		
multi-en	0.80	0.93	0.95	0.84	0.82
multi-de	0.71	0.85	0.85	0.78	0.75
multi-fr	0.73	0.87	0.83	0.79	0.79
multi-qu	-	-	0.70	0.64	-
			XLM-R		
multi-en	**0.87**	**0.95**	**0.96**	0.84	**0.84**
multi-de	**0.81**	**0.90**	**0.88**	0.78	**0.77**
multi-fr	**0.82**	**0.90**	**0.88**	0.79	**0.82**
multi-qu	-	-	0.72	0.64	-

Table 8: Results of NLP tasks for different models trained on four language pairs, each evaluated on Pair Sentence, Structured prediction, and QA tasks.