Multilingual CheXbert: Radiology Report Labeling in Spanish

Stanford CS224N Custom Project

Adriel Saporta
Department of Computer Science
Stanford University
asaporta@stanford.edu

Emily Ross
Department of Computer Science
Stanford University
emiross@stanford.edu

Abstract

Automatic label extraction from free-text radiology reports enables efficient and large-scale training of natural language processing models for the medical setting. The current state-of-the-art label-extraction model, CheXbert [1], has been shown to work well on English-language radiology reports, but has not yet been tested in the multilingual setting. In this work, we explore how well Multilingual BERT performs on Spanish-language radiology reports. We find that regardless of whether the model is finetuned on English reports or Spanish reports, Multilingual BERT offers no real performance gains over English BERT when evaluating on Spanish-language reports. Furthermore, we show that while finetuning on human-labeled reports is better than finetuning on automatically-labeled reports, finetuning first on automatically-labeled reports and then further finetuning on human-labeled reports offers the best results.

1 Key Information to include

- External collaborators (if you have any): Anuj Pareek (radiology resident for domain expertise)
- External mentor (if you have any): Pranav Rajpurkar (Stanford ML Group)
- Code: Our codebase can be found here: https://github.com/ASaporta/multilingual-CheXbert (we have given Akshay Smit and Pranav Rajpurkar access to it). Much of our codebase was adapted from the CheXbert repo: https://github.com/stanfordmlgroup/CheXbert. The following scripts are new: `evaluate.py`, `preprocess.py`, `split_data.py`, and `translate.py`.

2 Introduction

Chest X-rays, the most common radiological exam, are a crucial tool in healthcare, allowing healthcare practitioners to rule out, identify, or monitor the progress of various (often critical) health conditions. Chest X-rays are typically accompanied by textual summaries of key observations and findings made by the attending radiologist. These radiology reports represent a radiologist's professional interpretation of the chest X-ray image, including statements on which conditions are (1) unlikely (negative), (2) likely (positive), and (3) not to be ruled out as a possibility (uncertain). However, this information is encoded in language rather than a more structured format. Extracting structured labels from the reports could yield efficiency improvements in medical settings. For instance, healthcare practitioners could better prioritize patients if they could sort them by potential diagnosis severity. Structured labels are also needed in order to perform research (for instance, learning diagnoses from the image alone), and automatic methods would remove the burden from expert radiologists of manually labeling datasets that are sufficiently large.

In the past, researchers have explored automatic label extraction using a variety of approaches, including rule-based and neural methods. Rule-based methods failed to extract key insights obfuscated by complex natural language, and initial neural approaches, though successful, were highly dependent on large volumes of expert-annotated data for training without making use of existing labeler systems. More recently, a model called CheXbert [1], found great success by utilizing the scaling opportunity provided by existing rule-based labelers and also incorporating high-quality expert-labeled data.

CheXbert helps solve the limitations posed by small volumes of expert annotations in the medical domain. This problem is only exacerbated when we look to foreign languages, many of which are low-resource, and especially so when it comes to the medical domain. We explore the extent to which the methods used to develop CheXbert can be modified and applied in a multilingual setting. Specifically, we compare the performance on a positive label extraction task of multilingual BERT-base models [2] ("M-BERT") to the performance of the monolingual, English BERT-base model [3] ("EN-BERT") used by CheXbert. Our experiments entail finetuning the M-BERT model on various combinations of English- and Spanish-language, and automatically- and human-labeled reports, and then evaluating these models on human-labeled Spanish-language reports.

We find that regardless of whether the model is finetuned on English reports or Spanish reports, M-BERT offers no real performance gains over EN-BERT when evaluating on Spanish-language reports. Furthermore, we show that while finetuning on human-labeled reports is better than finetuning on automatically-labeled reports, finetuning first on automatically-labeled reports and then further finetuning on human-labeled reports offers the best results.

3 Related Work

Our research directly builds off of the efforts of the authors of CheXbert. CheXbert not only made efficient use of available resources, but also outperformed the previous state-of-the-art rule-based labeler with statistical significance. The CheXbert task was to classify each of 14 "observations" as blank, positive, negative, or uncertain. Observations included pathologies such as "Edema" or "Enlarged Cardiomediastinum", and non-pathological observations in the chest X-ray such as "Fracture" or "Support Devices". A "blank" label indicated that a particular observation wasn't mentioned in the summary. Negative and positive affirmed the respective absence or presence of a particular observation based on the summary. Uncertain meant the summary mentioned the observation without confirming or excluding the possibility of its presence. This 4-class paradigm played a critical role in our training process.

CheXbert was obtained by starting with a BERT model called BlueBERT, which was pretrained on a biomedical corpus [4]. This BERT model was then finetuned in two steps using CheXpert [5], a large publicly available dataset of chest X-rays. In the first step, the pretrained BlueBERT model was finetuned on labels automatically generated by the rule-based CheXpert labeler. In the second step, the model was finetuned on the CheXpert manual set, a subset of CheXpert that was labeled by expert radiologists, and augmented by backtranslation. Backtranslation is the process of translating a sequence into another language and then translating it back, exploiting the information loss inherent in imperfect translation models to yield semantically similar but syntactically different textual data. This was especially useful because of the limited amount of manually-labeled data available. Our approach differed in that we did not pretrain on a biomedical corpus, nor did we always finetune in two steps. We also did not employ backtranslation. Though we currently deviate from the CheXbert method in these ways, our experimental process was heavily informed by CheXbert and the conceptual similarities are not insignificant. Our future work will in part focus on thinning the gap between their methods and ours.

One of the ways CheXbert was evaluated was with an average of F1 scores calculated for each observation. According to this metric, CheXbert achieved statistically superior performance compared to CheXpert, the previous state-of-the-art labeler. CheXbert also achieved statistically superior performance on a per-observation level for 9 of the 14 observations. The authors analyze specific examples where CheXbert was able to identify positive, negative, and uncertain observation labels that CheXpert missed, demonstrating that the complexity of natural language stands to benefit from a more sophisticated approach than rule-based.

The success of CheXbert demonstrated that the overall approach of training on a combination of automatically- and manually-labeled data (and the use of backtranslation as a dataset augmentation technique) could be worth applying across other medical domains—and other languages. Where high-quality expert-labeled data is not so widely available or easy to obtain, training in conjunction with automatically-labeled data can yield promising results.

4 Approach

4.1 Task

The report labeling task is to take as input the *Impression* section of a free-text radiology report (which summarizes the key findings in a chest X-ray image) and output 14 labels for 14 conditions that could be seen on a chest X-ray. Each of the 14 labels is binary: a positive output indicates that the class is present according to the radiology report and a negative output indicates that the class is absent according to the radiology report.

4.2 Data

In this work, we use two large publicly available chest X-ray datasets. As our English-language dataset, we used MIMIC-CXR [6], which has 187,674 radiology reports that were automatically labeled for the 14 conditions of interest using the CheXpert labeler [5]. For 13 of the 14 conditions, the CheXpert labeler outputs four possible classes: position, negative, uncertain, and blank. For the "No Finding" condition, the CheXpert labeler outputs a binary positive or negative label. Uncertain indicates that the condition was mentioned in the report, but it is uncertain whether the condition is present in the chest X-ray image. Blank indicates that the condition was not mentioned at all in the report. We convert all blank labels in MIMIC to negative labels, so that the MIMIC dataset had only three labels: positive, negative, and uncertain.

As our Spanish-language dataset, we use PadChest [7], which has 84,170 radiology reports. Of those reports, 63,889 were automatically-labeled using a supervised method based on a recurrent neural network with attention mechanisms, and 20,281 were manually annotated by trained physicians. While MIMIC has only 14 labels, PadChest has 193 labels for radiographic findings or differential diagnoses. Therefore, in order to map each of the 193 PadChest labels to one or more of the CheXpert labels (or 'N/A', if there was no comparable label), we built a mapping with the help of Anuj Pareek, a radiology resident in the Stanford Machine Learning Group. Although the majority of the PadChest labels either have one corresponding CheXpert label or no corresponding label ("many-to-one" or "many-to-none"), a couple of the PadChest labels have two corresponding CheXpert labels ("one-to-many"). Each report in the PadChest dataset has a corresponding list of labels. We consider any condition in that list to be positive and any condition absent from that list to be negative.

We randomly shuffle and split the MIMIC dataset into a training set (185,174 reports) and a validation set (2,500 reports). We divide the PadChest dataset into automatically-labeled reports and human-labeled reports. We then randomly shuffle and further split each of those two groups of PadChest reports: the automatically-labeled reports are split into a training set (61,389 reports) and a validation set (2,500), and the human-labeled reports are split into a training set (15,281), a validation set (2,500), and a test set (2,500). See Table 1 the prevalence of each condition in each of our dataset splits (two splits for MIMIC and five splits for PadChest).

4.3 Model Architecture

For all of our models, we follow the approach taken by the authors of CheXbert and use a modification of the BERT-base architecture [3] that has 14 linear heads, one for each condition (see Figure 1). The text of each radiology report is tokenized, and the maximum number of tokens in each input sequence is capped at 512. The hidden state corresponding to the CLS token in the model's final layer is then fed as input to each of the 14 linear heads. Deviating slightly from CheXbert's approach, we have 13 of the 14 linear heads generate three class scores (positive, negative, and uncertain), and have the linear head for "No Finding" generate two class scores (positive and negative).

	MIMIC (English)		PadChest (Spanish)				
Condition	Train set auto-labeled	Val set auto-labeled	Train set auto-labeled	Val set auto-labeled	Train set human-labeled	Val set human-labeled	Test set human-labeled
Atelectasis	32,517 *(17.6%)*	457 *(18.3%)*	4,357 *(7.1%)*	161 *(6.4%)*	1,140 *(7.5%)*	165 *(6.6%)*	188 *(7.5%)*
Cardiomegaly	30,660 *(16.6%)*	376 *(15.0%)*	6,880 *(11.2%)*	272 *(10.9%)*	1,865 *(12.2%)*	292 *(11.7%)*	293 *(11.7%)*
Consolidation	8,365 *(4.5%)*	109 *(4.4%)*	1,333 *(2.2%)*	61 *(2.4%)*	178 *(1.2%)*	29 *(1.2%)*	27 *(1.1%)*
Edema	21,242 *(11.5%)*	277 *(11.1%)*	3,603 *(5.9%)*	145 *(5.8%)*	1,207 *(7.9%)*	186 *(7.4%)*	204 *(8.2%)*
Enlarged Cardiom.	4,932 *(2.7%)*	64 *(2.6%)*	2,113 *(3.4%)*	81 *(3.2%)*	583 *(3.8%)*	85 *(3.4%)*	92 *(3.7%)*
Fracture	3,220 *(1.7%)*	46 *(1.8%)*	2,797 *(4.6%)*	105 *(4.2%)*	865 *(5.7%)*	150 *(6.0%)*	142 *(5.7%)*
Lung Lesion	4,950 *(2.7%)*	46 *(1.8%)*	5,012 *(8.2%)*	215 *(8.6%)*	1,310 *(8.6%)*	219 *(8.8%)*	221 *(8.8%)*
Lung Opacity	37,158 *(20.1%)*	501 *(20.0%)*	11,285 *(18.4%)*	473 *(18.9%)*	2,734 *(17.9%)*	449 *(18.0%)*	442 *(17.7%)*
Pleural Effusion	40,099 *(21.7%)*	522 *(20.9%)*	5,407 *(8.8%)*	218 *(8.7%)*	900 *(5.9%)*	148 *(5.9%)*	168 *(6.7%)*
Pleural Other	1,478 *(0.8%)*	16 *(0.6%)*	2,050 *(3.3%)*	79 *(3.2%)*	947 *(6.2%)*	157 *(6.3%)*	160 *(6.4%)*
Pneumonia	12,991 *(7.0%)*	175 *(7.0%)*	3,856 *(6.3%)*	161 *(6.4%)*	816 *(5.3%)*	141 *(5.6%)*	136 *(5.4%)*
Pneumothorax	7,474 *(4.0%)*	94 *(3.8%)*	277 *(0.5%)*	11 *(0.4%)*	50 *(0.3%)*	4 *(0.2%)*	13 *(0.5%)*
Support Devices	44,797 *(24.2%)*	596 *(23.8%)*	10,560 *(17.2%)*	412 *(16.5%)*	1,637 *(10.7%)*	250 *(10.0%)*	298 *(11.9%)*
No Finding	68,983 *(37.3%)*	960 *(38.4%)*	25,285 *(41.2%)*	1,044 *(41.8%)*	6,647 *(43.5%)*	1,139 *(45.6%)*	1,086 *(43.4%)*
Total size of split	185,174 *(100.0%)*	2,500 *(100.0%)*	61,389 *(100.0%)*	2,500 *(100.0%)*	15,281 *(100.0%)*	2,500 *(100.0%)*	2,500 *(100.0%)*

Table 1: Number of positive examples for each condition in each of the dataset splits for MIMIC and PadChest. In parentheses are the percentage of positive examples over total size of the relevant dataset split (shown in the last row).

4.4 Training Details

Each of our BERT-base models is pretrained either on lower-cased English text (EN-BERT) or on cased text in the top 104 languages with the largest Wikipedias (Multilingual BERT, or M-BERT). EN-BERT contains 110M parameters and M-BERT contains 179M parameters.

Just as in CheXbert, all models are finetuned for eight epochs using cross-entropy loss and Adam optimization with a learning rate of 2×10^{-5}. The cross-entropy losses for each of the 14 conditions are added to produce the final loss. We finetune all layers of our models, including the embeddings. During training, we use a dropout layer before the 14 linear heads, and we periodically evaluate our model on the validation set and save the checkpoint with the highest performance according to the Cohen's Kappa statistic. Unless otherwise specified, all models are trained using one Tesla K80 GPU with a batch size of 18.

4.5 Evaluation

To label radiology reports, MIMIC uses three classes (positive, negative, and uncertain) and PadChest uses two classes (positive and negative). Therefore, we chose to generate binary predictions from each model's three class scores for 13 of the conditions (as mentioned, the model already outputs only two class scores for "No Finding"). For each of the 13 conditions with three class scores, we ignore the score for the class "uncertain" and assign to that condition whichever of the remaining two classes, positive or negative, has the higher score.

Since the goal of this study is to determine how our models perform on Spanish-language radiology reports, all models are evaluated on the PadChest human-labeled test set. We evaluate our models using the F1 score for each of the 14 conditions. We also compute a weighted F1 score, where each condition is weighted by the proportion of positive labels for that condition in the ground truth test set.

5 Experiments

5.1 Baselines

We ran two baselines, for which we finetuned EN-BERT on the English-language MIMIC dataset. During training, we evaluate our model on the validation set every 2,000 iterations. For our first baseline, we evaluated the model on the Spanish-language PadChest test set. For our second baseline, we first translated the test set into English using HuggingFace's MarianMT [8], and then evaluated the model on that translated test set.

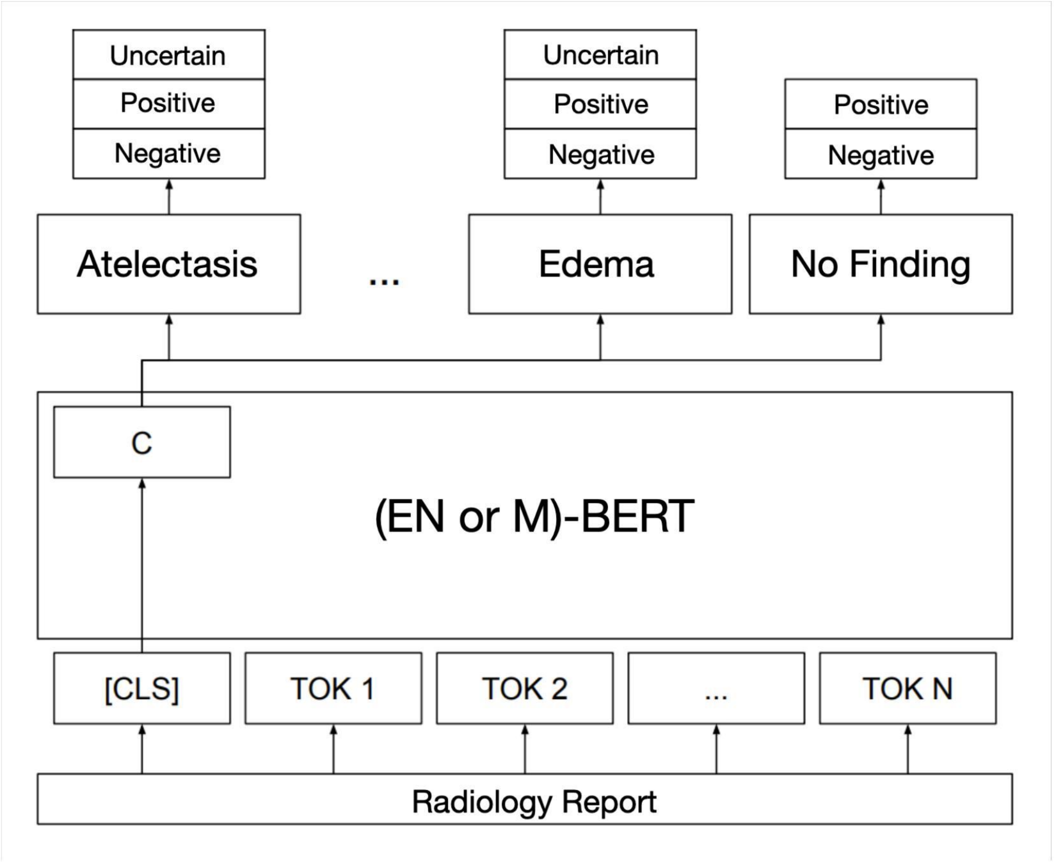


Figure 1: Following CheXbert's approach, all models use a modification of the BERT-base architecture with 14 linear heads.

Results As shown in Table 2, our first baseline, evaluated on the original test set in Spanish, achieves a weighted F1 score of 0.379. Our second baseline, evaluated on the translated test set in English, achieves a weighted F1 score of 0.492. This improvement in performance from our first baseline to our second baseline is likely due to the fact that EN-BERT is pretrained and finetuned on English datasets. Therefore, translating the test set's Spanish radiology reports into English before evaluation helped the model's performance significantly.

The F1 scores for Atelectasis are higher than for any other condition for both the first baseline (0.788) and the second baseline (0.950). Our first baseline obtains an F1 score of 0.000 for Fracture, Pleural Effusion, Pneumonia, and Pneumothorax. Our second baseline obtains an F1 score of 0.000 for Pleural Effusion. For all conditions with a baseline F1 score of 0.000 except Pneumonia, our models predict negative labels on all reports; for Pneumonia on our first baseline, our model predicts negative labels on all reports except for a single false positive prediction.

We hypothesize that Pneumonia is a particular difficult condition for EN-BERT given that 8.4% of the MIMIC radiology reports has an uncertain label for Pneumonia (significantly higher than for any other condition), and only 7.0% of the reports has a positive label for Pneumonia. Furthermore, the 14 conditions are a mix of radiographic findings (which are completely observable in a chest X-ray image) and differential diagnoses (which, according to the authors of PadChest, "are characterized by intrinsic uncertainty and a highly multidimensional context which is not included in the image"). The two differential diagnoses in our study are Pneumonia and Edema, on which neither our first baseline (Pneumonia 0.000; Edema 0.047) or our second baseline (Pneumonia 0.082; Edema 0.056) performs particularly well.

5.2 Experiments

We ran four experiments using M-BERT and one experiment using EN-BERT. For **M-BERT ft-en auto**, we finetune M-BERT on the English-language MIMIC dataset (whose labels were generated automatically using CheXpert). We evaluate our model on the validation set every 2,000 iterations. We use a Tesla P100-PCIE-16GB GPU instead of a Tesla K80 GPU.

For **M-BERT ft-sp auto**, we finetune M-BERT on the Spanish-language PadChest split of automatically-labeled reports. For **M-BERT ft-sp human**, we finetune M-BERT on the Spanish-language PadChest split of human-labeled reports. For **M-BERT ft-sp auto-human**, we initialize the model with the weights from M-BERT ft-sp auto and then finetune further on the PadChest split of human-labeled reports. During training for M-BERT ft-sp auto, M-BERT ft-sp human, and M-BERT ft-sp auto-human, we evaluate our model on the validation set every 200 iterations. See Figure 2 for the pipeline for M-BERT ft-sp experiments.

For **EN-BERT ft-sp auto-human**, we first finetune EN-BERT on the Spanish-language PadChest split of automatically-labeled reports, and then further finetune on the PadChest split of human-labeled reports. During both finetune steps, we evaluate our model on the validation set every 200 iterations and use a Tesla P100-PCIE-16GB GPU instead of a Tesla K80 GPU.

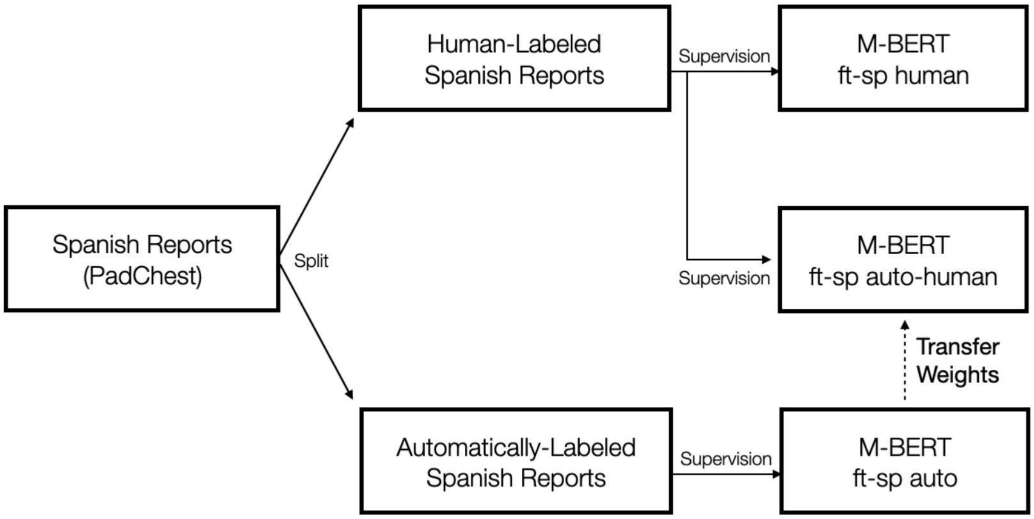


Figure 2: Pipeline for M-BERT ft-sp experiments.

Results As shown in Table 2, M-BERT ft-en auto obtains a weighted F1 score of 0.489, which is slightly worse than the weighted F1 score for our second baseline (0.492). While M-BERT ft-en auto is finetuned on English reports, we expected that the model would perform better than either of the two baseline given that it was pretrained on a multilingual dataset. However, it seems that using a multilingual BERT-base model gives no performance gains on a Spanish dataset than an English BERT-base model on a Spanish dataset naively translated into English before evaluation. We hypothesize that there are domain-specific terms that EN-BERT and M-BERT are both able to learn in English during finetuning, and that EN-BERT is able to recognized during evaluation after a naive translation, but that M-BERT isn't able to recognize in Spanish. Furthermore, just like the two baselines, M-BERT ft-en auto struggles on Pleural Effusion, for which it obtains an F1 score of 0.000.

Table 2 also shows the results for M-BERT ft-sp auto, M-BERT ft-sp human, and M-BERT ft-sp auto-human, all of which perform significantly better than the baselines and M-BERT ft-en auto across all 14 conditions. This indicates that finetuning M-BERT directly on Spanish-language reports significantly, and positively, impacts model performance on Spanish-language reports.

For 11 of the 14 conditions, M-BERT ft-sp human outperforms M-BERT ft-sp auto, which might be expected given that human labels are likely more accurate than automatically generated labels. That said, the automatically-labeled PadChest split is significantly larger than the human-labeled PadChest split, which could explain M-BERT ft-sp auto's better performance than M-BERT ft-sp human on Consolidation, Lung Lesion, and Pleural Effusion. Of the three M-BERT ft-sp models, M-BERT ft-sp auto-human performs the best on all conditions except Cardiomegaly and Pneumonia, and has the highest weighted F1 score (0.985), followed by M-BERT ft-sp human (0.979) and then M-BERT ft-sp auto (0.964).

Of all the conditions, the M-BERT ft-sp models perform the worst on Pneumothorax (auto 0.556; human 0.700; auto-human 0.870). We hypothesize that this is due to the low prevalence of Pneumothorax in the PadChest dataset: only 0.4% of the PadChest reports are positive for Pneumothorax. This suggests that prevalence is an important factor for model performance on a condition.

Finally, Table 2 shows the results of EN-BERT ft-sp auto-human. Of all the experiments we run, EN-BERT ft-sp auto-human has the highest weighted F1 score (0.986), and performs the best on 8 of the 14 conditions. Since our second baseline, even using a subpar translation model, has a higher weighted F1 score than M-BERT ft-en auto, and since the performances of M-BERT ft-sp auto-human and EN-BERT ft-sp auto-human are fairly comparable, it seems that M-BERT offers no real performance gains over EN-BERT, regardless of whether the model is finetuned on English reports or Spanish reports.

	EN-BERT	EN-BERT	M-BERT	M-BERT	M-BERT	M-BERT	EN-BERT
	ft-en auto	ft-en auto	ft-en auto	ft-sp auto	ft-sp human	ft-sp	ft-sp
	test set: orig	test set: trans				auto-human	auto-human
Condition	*(Baseline 1)*	*(Baseline 2)*					
Atelectasis	0.788	0.950	0.966	0.981	0.992	**0.997**	**0.997**
Cardiomegaly	0.683	0.673	0.685	0.920	**0.988**	0.978	0.980
Consolidation	0.432	0.333	0.303	0.863	0.848	0.962	**0.981**
Edema	0.047	0.056	0.092	0.981	**0.990**	0.990	**0.993**
Enlarged Cardiom.	0.250	0.210	0.328	0.885	0.889	**0.906**	0.902
Fracture	0.000	0.584	0.737	0.982	0.983	**0.986**	0.983
Lung Lesion	0.347	0.416	0.633	0.964	0.943	**0.982**	0.973
Lung Opacity	0.058	0.223	0.181	0.916	0.969	0.970	**0.974**
Pleural Effusion	0.000	0.000	0.000	0.988	0.985	0.988	**0.994**
Pleural Other	0.056	0.140	0.111	0.978	0.981	**0.991**	0.987
Pneumonia	0.000	0.082	0.043	0.967	0.974	0.960	**0.975**
Pneumothorax	0.000	0.375	0.000	0.556	0.700	**0.870**	**0.870**
Support Devices	0.013	0.443	0.037	0.987	0.993	**1.000**	0.997
No Finding	0.743	0.780	0.828	0.988	0.994	0.996	**0.997**
Weighted Average	0.379	0.492	0.489	0.964	0.979	0.985	**0.986**

Table 2: The F1 scores for each condition and the weighted F1 scores across all conditions for all six experiments, including baseline. Underlined values indicate best F1 score between the two baseline experiments, and bolded values indicate the best F1 score for each condition across all experiments.

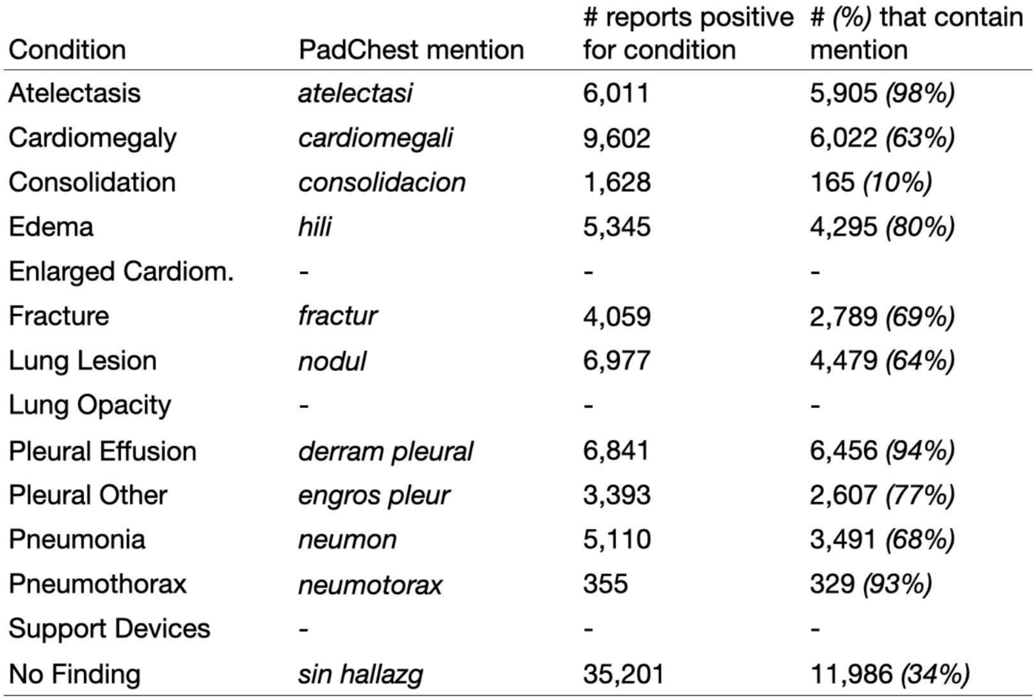
6 Analysis

We analyze why our second baseline, EN-BERT ft-en auto evaluated on the translated test set, performs more poorly than expected and find that the MarianMT translations are noticeably lacking. For example, a PadChest report that is positive for Lung Opacity and Pleural Other reads, *"cambi pulmonar cronic . engros pleuroparenquimat biapical . sign atrap aere con aplan diafragmat ."* The English translation produced by MarianMT is, *"cambi pulmonary cronic. engros pleuroparenquimat biapical. sign atrap aere con aplan diaphragm."* We see that there are abbreviations of fairly basic words in Spanish that MarianMT fails to translate at all. For example, *cambi* and *cronic* are abbreviations that likely mean *cambio* ("change" in English) and *cronico* ("chronic" in English), respectively. More surprisingly, even full, common Spanish words such as *con*, which means "with" in English, do not seem to be translated at all by MarianMT.

Given this limitation of MarianMT, it is not surprising that the translation model fails to recognize medical abbreviations such as *pleuroparenquimat*, which is short for *pleuroparenquimatos* meaning "pleuroparenchymal" in English. This inability to translate Spanish abbreviations, medical or otherwise, is apparent across reports in the test set and across conditions, and is likely a large reason that our second baseline performs so poorly.

This limitation of MarianMT does not explain, though, why both baseline models seem to perform fairly well on Atelectasis. To investigate this further, we choose the Spanish word or prefix most commonly associated with 11 of the conditions in the PadChest dataset and count the number of times that that word or prefix is mentioned in reports that are positive for each of the conditions (see Figure 3). For example, *neumon* is the Spanish prefix most commonly used in the PadChest reports to refer to Pneumonia.

Three conditions–Enlarged Cardiomediastinum, Lung Opacity, and Support Devices–do not have a single most common word or prefix associated with them, so we exclude them from the list. For example, a positive label for Enlarged Cardiomediastinum often co-occurs with the Spanish abbreviation for "mediastinum", which is *mediastin*. Enlarged Cardiomediastinum is also often indicated by terms such as *aument siluet cardiac* (meaning "increased cardiac silhouette" in English) and *elongacion aort* (meaning "aortic elongation" in English). Even for the 11 conditions that do seem to have a natural Spanish translation, they are often referred to by multiple names. For example, Lung Lesion is indicated by *nodul* ("nodule"), only slightly more often than it is indicated by *granulom calcific* ("calcified granuloma") or *masa* ("mass"). Because Atelectasis is consistently referred to as *atelectasi* in PadChest, it is unsurprising that the M-BERT ft-sp models perform well on it. Furthermore, since *atelectasi* is so close to the English "atelectasis," this could explain why the baseline models and M-BERT ft-en auto are able to still perform well on this condition. That said, it is especially surprising that the baseline models and M-BERT ft-en perform so poorly on Pleural Effusion, especially given that the terms in English and Spanish share the word "pleural".



Condition	PadChest mention	# reports positive for condition	# *(%)* that contain mention
Atelectasis	*atelectasi*	6,011	5,905 *(98%)*
Cardiomegaly	*cardiomegali*	9,602	6,022 *(63%)*
Consolidation	*consolidacion*	1,628	165 *(10%)*
Edema	*hili*	5,345	4,295 *(80%)*
Enlarged Cardiom.	-	-	-
Fracture	*fractur*	4,059	2,789 *(69%)*
Lung Lesion	*nodul*	6,977	4,479 *(64%)*
Lung Opacity	-	-	-
Pleural Effusion	*derram pleural*	6,841	6,456 *(94%)*
Pleural Other	*engros pleur*	3,393	2,607 *(77%)*
Pneumonia	*neumon*	5,110	3,491 *(68%)*
Pneumothorax	*neumotorax*	355	329 *(93%)*
Support Devices	-	-	-
No Finding	*sin hallazg*	35,201	11,986 *(34%)*

Figure 3: For 11 of the 14 conditions, we choose the Spanish word or prefix most commonly associated with that condition. We then count the number of times that that word or prefix is mentioned in the PadChest reports that have positive labels for that condition.

7 Conclusion

In this study, we extend the work of CheXbert to explore the relative performance of EN-BERT and M-BERT on the task of radiology report labeling. We reach three important conclusions. First, we find that regardless of whether the model is finetuned on English reports or Spanish reports, Multilingual BERT offers no real performance gains over English BERT when evaluating on Spanish-language reports. Second, while finetuning on human-labeled reports is better than finetuning on automatically-labeled reports (even if the dataset of human-labeled reports is smaller), it is best to first finetune on automatically-labeled reports and then further finetune on human-labeled reports. Third, keeping the BERT model type fixed, finetuning on Spanish reports offers significant gains when evaluating on Spanish-language reports than finetuning on English reports.

There are several limitations to our work. First, due to some ambiguities, our mapping from 193 PadChest to 14 CheXpert labels is somewhat subjective. Future work should incorporate consensus on this mapping among several radiologists. Second, it will be important to investigate further why both baselines and M-BERT ft-en auto perform so poorly on Pleural Effusion, a condition that past research suggests is easier to label than other conditions. Third, neither of the authors of this paper is a domain expert, and future work should work closely with a Spanish-speaking radiologist for a more thorough qualitative analysis of potentially problematic abbreviations in our dataset. Fourth, the translation model that we used for our second baseline, EN-BERT ft-en auto evaluated on the translated test set, was noticeably faulty. It would be interesting to explore how much better the second baseline would perform if we were to use a stronger translation model than MarianMT.

Our hope is that this study serves as the basis for future work that explores radiology report labeling in the multilingual setting.

References

[1] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, November 2020. Association for Computational Linguistics.

[2] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert?, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[4] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics.

[5] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, Jul. 2019.

[6] Alistair E W Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[7] Aurelia Bustos, Antonio Pertusa, Maria Salinas, and Maria de la Iglesia Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 08 2020.

[8] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.