Improving Code-Switched NLP: Fine-Tuning Multilingual Models on Monolingual, Bilingual, and Mixed Datasets

Majd Aldaye
ma798@cornell.edu

Pun Chaixanien
sc2343@cornell.edu

Nicholas Chernogor
nac86@cornell.edu

Kidus Zegeye
kmz25@cornell.edu

Abstract

The ability of language models to capture code-switching (CS) is crucial for processing real-world data from multilingual speakers. In this paper, we examine the intersection between multilingual pre-trained language models (PLMs) and code-switching through replicating two CS detection tasks: sentence classification and token-level language identification, and our results confirm previous findings on these tasks. Following this, we explore the effect of fine-tuning multilingual BERT on sentiment analysis and part-of-speech tagging using datasets of different languages. We found that the bilingual model performed best overall, while the English and CS-trained models performed best on CS data. We further find that accuracy for POS tagging is high overall, especially for the model fine-tuned on bilingual data, but CS test data is best captured by a model fine-tuned on CS data. This highlights the need to consider CS-specific tasks in development of multilingual models.

Introduction

Code-switching (CS) is a linguistic phenomenon common in multilingual communities; it occurs when speakers alternate between different languages within a conversation or sentence. In order to process linguistic information in real-world contexts—where multilingual speakers exist and employ CS in their speech and writing—ensuring that language models are able to handle CS data is therefore crucial. This leads us to examine the intersection between multilingual pre-trained language models (PLMs) and code-switching, replicating another paper such as to examine the question of whether their multilingual functionality sufficiently equips them for the task of detecting CS data on the sentence level and correctly identifying the language on a token level.

Extending from this, we noticed that PLMs are often used in the context of fine-tuning for a particular task, rather than simply being employed off-the-shelf for their pre-trained functionality. We appreciate that these PLMs have been trained on vast amounts of multilingual data and multilingual capability is something we want to retain. Therefore, we question how fine-tuning on a particular language or combination of languages for a particular task would affect multilingual model performance.

This question has many implications for less-common languages that lack labeled data for downstream tasks: if we find that fine-tuning a multilingual model on a task in one language also allows it to perform the task on its other languages, this may reduce the need for gathering and labeling data and fine-tuning models for low-resource languages. Furthermore, this could address the question of differences between bilingual and CS data, informing us of the value of gathering CS data in particular.

This paper examines fine-tuning in the context of two particular tasks, sentiment analysis and part-of-speech tagging, using datasets in different languages, in order to explore the effect on multilingual capability in multilingual PLMs and whether the language we fine-tune the model on matters. Our code for our work can be found here: https://github.com/Pun2341/cl2-code-switching-plm

Related Works

Probing Language Models for Code-Switched Text

Pre-trained language models (PLMs) such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have shown remarkable capabilities in encoding cross-lingual information. Probing techniques, which train auxiliary classifiers to examine the linguistic properties encoded in model representations, have been widely used to explore morphological, syntactic, and semantic generalization (Tenney et al., 2019; Hewitt and Manning, 2019). In the context of CS, studies like Santy et al. (2021) have demonstrated that PLMs encode syntactic structure and semantic meaning of mixed-language text, with varying effectiveness depending on data quality and task complexity.

Laureano De Leon et al. (2024) explores PLMs' ability to generalize on Spanglish CS text. The paper evaluates PLMs on three key dimensions:

1. Detection: Differentiating between monolingual and CS sentences at both sentence and token levels through Language Identification (LID) experiments.
2. Syntax: Comparing the grammatical structure of CS text to its monolingual translations using dependency parsing and Graph Edit Distance (GED).
3. Semantics: Evaluating semantic consistency between CS and monolingual representations using Semantic Text Similarity (STS) tasks.

Their findings highlight that PLMs generalize effectively to CS text without explicit training, particularly for syntactic and semantic representations. However, the performance degrades when using synthetic CS data, underscoring the importance of naturalistic examples for evaluation.

Existing Benchmarks and Datasets

The lack of standardized benchmarks for CS evaluation has led to the development of datasets such as LinCE (Aguilar et al., 2020) which include labeled CS text for tasks like LID, POS tagging, machine translation (CALCS 2016 dataset, Molina et al., 2016), and sentiment analysis (SentiMix, Patwa et al., 2020). However, challenges remain in generating synthetic CS text that reflects naturalistic language patterns. Methods such as token replacement and noun-phrase substitution (Krishnan et al., 2021; Salaam et al., 2022) often produce well-formed sentences but fail to capture the nuanced structure of real-world CS data.

Replication

To validate and build upon the findings of Laureano De Leon et al. (2024), we replicate their probing experiments, focusing on the detection tasks: code-switching identification on the sentence level and Language Identification (LID) on the token level.

Models

We evaluated the performance of two pre-trained models:

- mBERT (bert-base-multilingual-cased)
- DistilBERT (distilbert-base-uncased)

Given hardware constraints, mBERT was tested with 4-bit quantization on GPU to improve memory efficiency, while DistilBERT was used for CPU-based experiments. All implementations used the HuggingFace Transformers library.

Target Layers

To replicate layer-wise probing from the original study, we extracted representations from the 3rd, 6th, and 11th layers. These layers provide insight into the progression of language-specific information as it propagates through the model.

Datasets

We used two code-switched datasets, as in Laureano De Leon et al. (2024):

1. SentiMix 2020 (Patwa et al., 2020)
2. CALCS 2016 (Molina et al., 2016)
3. ProfNER 2021: a monolingual dataset of tweets in Spanish and English (Miranda-Escalada et al., 2021)

Results & Analysis

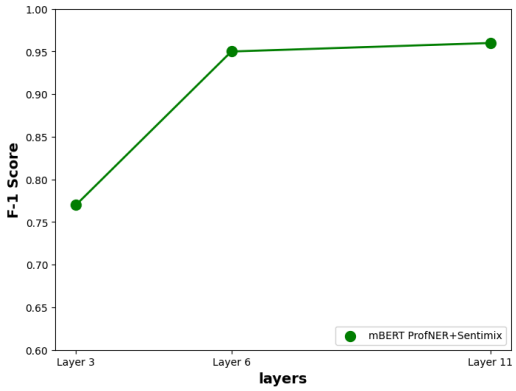

Figure 1: Layer-wise F1 Scores for mBERT on the SentiMix dataset.

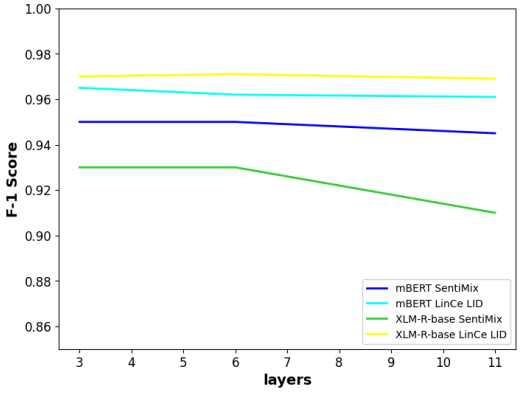

Figure 2: Layer-wise F1 Scores for mBERT and XLM-R-base on SentiMix and LinCE datasets.

Our replication confirms that pre-trained multilingual models like mBERT and XLM-R-base effectively encode language identity across layers. XLM-R-base demonstrates superior performance, particularly for cleaner datasets like LinCE. These findings align with Laureano De Leon et al. (2024) and highlight the robustness of PLMs for detection tasks.

Extension Methodology

Datasets

For the fine-tuning tasks, we used datasets that contained labels for sentiment data and POS tagging for sentences in English, Spanish, and English-Spanish code-switched. The code-switched datasets are hosted by LinCE, where the POS tagged data was sourced from AlGhamdi et al. (2016) and the sentiment data was sourced from Patwa et al. 2020. We sourced the English and Spanish sentiment datasets from an aggregate of datasets hosted by Qiang 2022. We sourced the English POS data from Liu et al. 2018 and the Spanish POS data from Taulé 2008.

For sentiment analysis, each training dataset consists of 1800 samples and the test sets consist of 900 samples. The reason for such a small sample size is due to the constraint of the smallest dataset. Each entry in the dataset was labeled with 'positive', 'negative', or 'neutral' sentiment classes.

For POS tagging, each training dataset consists of 1394 sentences and each test set consists of 1200 sentences. While more data was available in the code-switched and Spanish datasets, sample sizes were abridged to the length of the English dataset. Prior to training and testing, slight modifications were made to the datasets such as to ensure labels were consistent across datasets. The original code-switched dataset contained the label 'UNK' for unknown words, which was not present in the Spanish or English data, so these labels were instead mapped to 'X', which all three datasets did contain as a label for words that could not be given an appropriate POS tag. The Spanish and English datasets also contained 'SYM' as a label for symbols such as emojis, emoticons, percent signs, currency symbols, etc. Since the code-switched dataset did not contain this label, all 'SYM' labels were also mapped to 'X'. Furthermore, constituting conjunctions were labeled as 'CONJ' in the code-switched dataset but 'CCONJ' in the Spanish and English data, so these were all mapped to 'CONJ'. Finally, the Spanish dataset expressed multimorphemic words in the following format: "del _ / de ADP / el DET," while in the code-switched dataset the format was "de ADP / el DET," so the Spanish data was cleaned to remove words tagged with "_", as well as removing dropped pronouns which were tagged as "_ PRON". This resulted in consistent labelling schemes across the datasets, allowing for fair comparisons in performance.

Models and Evaluation

Our main purpose is to look into how fine-tuning with different languages affects model performance. The model we decided to use is the BERT Multilingual Cased (Devlin et al., 2019). This ensures that the model already has multilingual capability and allows us to determine how fine-tuning for a particular task in a particular language will influence the models' ability to perform that task in different languages. To allow for a fair comparison, we decided to split into 2 different tasks: sentiment analysis, which classifies an entire input, as well as part-of-speech tagging which classifies every individual token. Then, for each task, we fine-tune four different models using different data: code-switched data, English data, Spanish data, and bilingual data. Bilingual data simply consists of a random 50% of English data and 50% of Spanish data. The four models use the same number of samples and the same parameters. Then, for each model, we test its ability to perform that task on different test sets: code-switched test set, English test set, and Spanish test set. For both tasks, we use a batch size of 32, we run training over 10 epochs at a learning rate of 1e-4.

Results

Sentiment Analysis

Figure 3 shows the accuracy scores for each model compared to how it performed on different test sets. We notice that for code-switched predictions, using code-switched data or English data works the best. For monolingual English and Spanish data, we find that using a combination of English and Spanish perform the best, even out performing monolingual training. Testing on a combination of all test sets, we find that the bilingual model performs the best.

	Test set				
Model		**CS**	**English**	**Spanish**	**All**
	CS	**0.56875**	0.33	0.33125	0.381875
	English	**0.56875**	0.33125	0.33125	0.3821875
	Spanish	0.15	0.34125	0.33875	0.2946875
	Bilingual	0.46625	**0.55875**	**0.47**	**0.50125**

Figure 3: Accuracy Scores for Sentiment Analysis

POS Tagging

Testing the fine-tuned models on the POS tagging task produced the results seen in Figure 4. We see that, for each test set, the model fine-tuned on the corresponding language type performed best. When combining all three of the test sets (see the 'All' column), the bilingual model has the best overall performance.

	Test set				
Model		**CS**	**English**	**Spanish**	**All**
	CS	**0.94**	0.66	0.91	0.85
	English	0.81	**0.93**	0.87	0.88
	Spanish	0.79	0.66	**1.00**	0.87
	Bilingual	0.84	0.91	0.99	**0.95**

Figure 4: Accuracy scores for part-of-speech tagging

Discussion

Sentiment Analysis

The method used to score was simply to compare what the model outputted with the ground truth. However, we also considered a sort of loss scheme where we examine the likelihoods attached to each class (negative, neutral, or positive). The results of that scheme turned out to be unhelpful which is why we decided to use only accuracy. Seeing as there are 3 possible classes, ⅓ can be seen as random guessing. There is only one instance that falls below this, which suggests some issue with the model. All

in all, however, we notice that the Bilingual model performs very highly the most consistently, indicating to us that this may be the most effective way to fine-tune for sentiment analysis if we would like to keep the multilingual capabilities in mind.

When testing on a monolingual test set, we note that accuracy scores are around random guessing for all models other than the bilingual one. The only outlier for these models is their performance on the codeswitched test set. Here, other than the Spanish model, the others perform very well with the English and code-switched fine-tuned models performing best. Regarding the Spanish model, it comes as a large surprise to see that the model performs extremely poorly. This could indicate difficulties in training as well as issues in the data itself. One possible idea regarding this is about what language the code-switched data is grounded in. Oftentimes, people who code-switch are likely to use a language as their main one and another to fill in certain words to help bridge gaps in meaning. As such, if the data were to be in English with certain words in Spanish or the other way around, this factor could largely affect performance.

POS Tagging

Across all permutations of model and dataset, we note that accuracy is relatively high. Given the 17 possible tags, we assume that a random baseline would achieve accuracy of only around 0.06, while even the lowest accuracies in our results far surpass this. We note that the English-fine-tuned model reaches an accuracy of 0.87 on the Spanish test set (though we don't see the same vice versa, which is likely attributable to the nature of the English dataset as will be discussed in the Limitations section). This lends credence to the idea that fine-tuning a multilingual model on POS tagging does not necessarily require providing it data from each of the languages it was pre-trained on in order to reach an adequate level of performance.

However, we do still find that the models fine-tuned on a particular training set performed best on the corresponding test set, leading to the highest accuracies being present in the diagonals of the Results table: 0.94 for CS-fine-tuned on CS data, 0.93 for English-fine-tuned on English data, and 1.00 for Spanish-fine-tuned on Spanish data. This aligns with intuition: we would not expect a model that has been fine-tuned on Spanish data to outperform a model fine-tuned on English data when testing on English.

Another relevant observation is that the bilingual-fine-tuned model performs the best overall, with 0.95 accuracy on the combination of all three test sets. It is able to achieve comparable accuracy on both the English and Spanish test sets (0.91 and 0.99) as those respective monolingual-fine-tuned models are able to achieve (0.93 and 1.00), despite only having seen half of the data from each monolingual training set during fine-tuning. Thus we note that it is effective to fine-tune on multiple languages, even if it means having less data for each language, without significantly reducing performance.

At the same time, we note that bilingual-fine-tuned model's lowest performance is on the code-switched data (0.84), significantly lower than the 0.94 accuracy that the CS-fine-tuned model achieves on CS data, suggesting that code-switching cannot be fully captured by a multilingual model that has not specifically seen CS data. This motivates the creation and use of more CS data when testing the performance of multilingual models, such as to ensure models are able to capture this very common phenomenon for multilingual speakers.

Limitations

The limitations of the sentiment analysis fine-tuning largely comes from the dataset used. The training data consists of only 1800 samples, a relatively small dataset, which could be limiting the generalizability of the results. Furthermore, the data was in the form of tweets, which present unique challenges such as contamination from links, emojis, and other non-standard textual elements. The length of tweets also makes it difficult for models to capture nuanced sentiments effectively. Additionally, while the training process was capped at 10 epochs to avoid overfitting, increasing the number of epochs could have allowed the model to better utilize the training data.

For the POS datasets, several limitations should be considered which may have influenced the results achieved. Firstly, each dataset is somewhat different in nature as they come from different sources: the CS data appears to be sentences from spoken conversations, while the English data is sourced from Tweets, and the Spanish data appears to include primarily formal written sentences. Following this, each dataset also has different lengths of sentences: many of the spoken sentences in the CS dataset are very short, while the formally-written sentences in the Spanish dataset are lengthier. This means that, while the models were trained on the same number of sentences, they saw different quantities of data. These factors may have contributed to the low performance of the Spanish- and CS-fine-tuned models on the English dataset, as they have not been fine-tuned on Tweets, and the high performances of the Spanish- and bilingual-fine-tuned models on the Spanish data, as they have seen long Spanish sentences with ample labels during fine-tuning.

Future Work

Extending our analysis on these models to other tasks beyond sentiment analysis and POS tagging is a natural next step for this project. This could include named entity recognition (NER), language identification, and machine translation. These tasks could glean more information on what kind of tasks generalize over different fine-tuned languages better. It would especially be interesting to further investigate machine translation with CS sentences, as the syntactical rules for the CS words or phrases can be undefined given the sentence is in a different language already.

We also are interested in experimenting with other PLMs. Currently we only use mBert for our main experiment, but we could extend to using DistillBERT, RoBERTa, and other models in various sizes to see how they correlate with the task. If smaller models like DistillBERT have similar performance to mBERT, it could save on training time and overall resource footprint.

There is also the area of other language interactions. We only experimented with English and Spanish, but datasets existed for other code-switched language pairs. Because English and Spanish are the most common languages in the United States, the models could have more experience with those languages. It would also be useful to see how languages with various syntactic structures, language families, and resources compare to each other in terms of trainability.

Bibliographical References

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wen zek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pages 8440–8451, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Hewitt and Christopher D Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In NAACL.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, Ellie Pavlick, and Google AI Language. 2019b. What do you learn from context? Probing for Sentence Structure in Contextualized Word Representations. In *International Conference on Learning Representations.*

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix * How does Code-Mixing interact with Multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP,* pages 111–121.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the Second Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the Second Workshop on Computational Approaches to Code-Switching*, pages 40–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual Code-Switching for Zero-Shot Cross Lingual Intent Prediction and Slot Filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cesa Salaam, Franck Dernoncourt, Trung Bui, Danda Rawat, and Seunghyun Yoon. 2022. Offensive Content Detection Via Synthetic Code Switched Text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6617–6624.

Language Resources References

Chen, Shuguang and Aguilar, Gustavo and Srinivasan, Anirudh and Diab, Mona and Solorio, Thamar. 2022. *CALCS 2021 Shared Task: Machine Translation for Code-Switched Data*. PID https://ritual.uh.edu/lince/datasets.

Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of Speech Tagging for Code Switched Data. In Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 98–107, Austin, Texas. Association for Computational Linguistics.

Patwa, Parth and Aguilar, Gustavo and Kar, Sudipta and Pandey, Suraj and PYKL, Srini vas and Gambäck, Björn and Chakraborty, Tan moy and Solorio, Thamar and Das, Amitava. 2020. *SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets.* Association for Computational Linguistics. PID https://ritual.uh.edu/lince/datasets.

Qiang, Yong. 2022. *multilingual-sentiment-datasets*. GitHub repository. PID https://github.com/tyqiangz/multilingual-sentiment-datasets

Taulé, M., M.A. Martí, M. Recasens (2008) 'Ancora: Multilevel Annotated Corpora for Catalan and Spanish', Proceedings of 6th International Conference on Language Resources and Evaluation. Marrakesh (Morocco).

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing Tweets into Universal Dependencies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.