Exploring POS & LID Tagging in Code-Switched Text

Rocky Klopfenstein, Christian Manzi, Gretta Ineza
Amherst College
{rklopfenstein27, cmanzi25, gineza26}@amherst.edu

Abstract

Code-switching, changing from one language to another, often results in grammatical or structural inconsistencies. These inconsistencies pose as difficult challenges for NLP systems, which may view such inputs as misspelled or 'noisy', resulting in poor performance. To address this problem, we explored approaches to improve Part-of-Speech (POS) and Language Identification (LID) tagging in code-switched text. We evaluated the performance of fine-tuning transformer-based multilingual models under different training setups, such as single task (ST), sequential task, and joint multitask learning (JMT). Though not conclusive, our results point to sequential learning and loss-weighted training as promising methods for improving tagging accuracy. Ultimately, our goal is to refine downstream systems, such as translational technology and LLMs, thus supporting greater linguistic equity in natural language processing.

1 Introduction

In many multilingual communities, code-switching, the practice of alternating between languages within a single utterance or conversation, is a natural and common form of communication. It's becoming so common that some language pairs have been given pseudo-names, such as "Spanglish" (Spanish & English) and "Chinglish" (Chinese & English). Yet, most state-of-the-art natural language processing (NLP) systems, namely LLMs, were not designed to handle such linguistic complexity. This can be partially attributed to the lack of training data surrounding code-switching, in comparison to dominant training languages such as English. For example, (Vyas et al., 2014) shows that models trained on monolingual data perform significantly worse on code-mixed Hindi-English text. Moreover, foundational tasks like Part-of-Speech (POS) and Language Identification (LID) tagging, which serve as building blocks for downstream applications such as parsing and machine translation, often suffer from degraded performance when applied to code-switched text, due to similar reasons stated above. This performance gap can, and threatens to, magnify digital inequalities, especially for first-generation students and speakers of minority and under-resourced languages who frequently engage in code-switching in informal and institutional settings alike.

Furthermore, as AI systems, including LLMs, become integrated into essential parts of our lives, including education, healthcare, and government services, linguistic bias in NLP threatens to widen existing social divides. Entire communities that engage in code-switching, often from underrepresented or marginalized communities, may find these implemented systems to be less responsive, less accurate, or even outright inaccessible. For instance, important automated calls, such as with government programs or healthcare companies, may misinterpret user inputs, and automated translation systems used in schools or public transit may mistranslate code-switched sentences. Not only do these errors reduce the utility of NLP systems as a whole, and potentially endanger users when deployed in important settings, but they also risk reinforcing harmful stereotypes about "correct" or "intelligent" ways of speaking.

To address the rising linguistic divide, it is necessary to first overcome a key challenge in processing code-switched text, which is that it violates assumptions made by monolingual and even multilingual models trained on clean, single-language datasets (Solorio and Liu, 2008). These models often mislabel words from minority languages, struggle with ambiguous or shared vocabulary, and overfit to dominant languages. These problems are only exacerbated by the lack of annotated, code-switched datasets (Aguilar et al., 2020). In response to these difficulties, we explore whether

different fine-tuning strategies can improve POS and LID tagging in code-switched tweets. Since both POS and LID tags provide crucial linguistic information, such as identifying the language and function of each word, they are foundational for understanding code-switched inputs, and overcoming key limitations facing current LLMs.

To be more specific, we believe that Part-of-Speech (POS) and Language Identification (LID) tagging are deeply related tasks in code-switched text and could be mutually beneficial in a fine-tuning setting (Soto and Hirschberg, 2018). Language identity provides important context for POS tagging, especially when a word may carry different grammatical roles across languages. For instance, in Spanish, the word "su" acts as a *possessive adjective*, whereas in Italian, it's used as a *preposition*. In this case, language identification directly supports POS tagging.

Conversely, specific POS patterns can help clarify language identity by revealing syntactic patterns that are language-specific. The simplest example is "no". The word "no" typically acts as an interjection (INTJ) in *English*, but is normally an adverb (ADV) in *Spanish*. In both cases, an otherwise tricky classification decision is made much clearer by either POS or LID tags. By training various multilingual models to perform both tasks, whether at the same time or sequentially, we aim to improve the accuracy and reliability of both POS and LID tagging (Khanuja et al., 2020). This coupling is especially valuable in noisy, low-resource settings like code-switched text, where one task can provide stability and signal for the other.

To test this hypothesis, we focused on the LinCE CALCS English-Spanish dataset (Aguilar et al., 2020), an annotated dataset of English–Spanish code-switched tweets, and evaluated the performance of three multilingual transformer models, XLM-R, RemBERT, and mBERT. We fine-tuned these models under four training paradigms: single-task, sequential learning, joint multitask learning, and loss-weighted multitask learning. Single-task training served as a baseline for comparisons against the joint approaches, and mBERT served as a baseline model. Using F1 score, accuracy, and qualitative analysis, our results show that multitask training strategies, particularly sequential learning and loss-weighted multitask learning, can modestly improve tagging performance in code-switched settings. While the improvements are limited and likely context-dependent, the trends are consistent and highlight the potential of multitask training strategies for POS and LID tagging. We hope to create a foundation that can be further expanded upon in future works, ultimately contributing to linguistic equity in natural language processing.

2 Related Work

Early approaches to tagging code-switched text often relied on combining monolingual tags or using heuristic-based methods. For example, (Solorio and Liu, 2008) proposed a system that combined POS tags from English and Spanish, achieving notable improvements over individual, monolingual tagging. Similarly, Barman et al. (2014) used classifiers with n-gram and contextual features for LID tagging in Bengali-Hindi-English datasets, once again attaining improvements over monolingual taggers.

Another example is from AlGhamdi et al. (2016), which explored various methods for POS tagging in code-switched text, including combinations of monolingual data. Their experiments on Spanish-English and Modern Standard Arabic-Egyptian Arabic datasets found that models trained directly on code-switched data outperformed combinations of monolingual data. This reinforces the somewhat intuitive, but still important, concept that code-switched specific training data, rather than monolingual data, is vital to fine-tuning a model's performance.

Furthermore, there has been interesting work around the creation of benchmarks for evaluating code-switched models. Khanuja et al. (2020), a research team from Microsoft Research India, created GLUECoS, an evaluation benchmark for code-switched NLP tasks, including POS and LID tagging. They also evaluated other tasks such as Named Entity Recognition (NER), Sentiment Analysis (SA), Question Answering (QA), and Natural Language Inference (NLI). Their work not only provided an important and standard benchmark for evaluating code-switched models on various tasks, but also found that multilingual models fine-tuned on code-switched data perform better than those trained solely on monolingual data, once again reinforcing the value of code-switched specific training data.

Finally, prior work has also set a precedent for POS and LID joint training in code-switched text. Soto and Hirschberg (2018) introduced a bidirectional Long Short-Term Memory model, trained

on the Miami-Bangor corpus, which achieved a POS tagging accuracy of 96.34% and a LID accuracy of 98.78%. Their model significantly outperformed previous state-of-the-art taggers by effectively leveraging joint training for POS and LID tasks, and also by having the ability to handle intra-sentential code-switching. For reference, intra-sentential code-switching means switching between languages within the same sentence.

In all, these studies indicate that models specifically trained on code-switched training data, specifically those utilizing joint training methods, achieve better performance in POS and LID tagging. Rather than introducing a new neural network architecture, our work builds on this foundation by exploring novel fine-tuning methods, such as sequential learning and loss-weighted multitask training, to further refine tagging accuracy in code-switched text. In doing so, we hope to bring together the insights from all these previous efforts in a unified and practical approach. Given more time, we would love to submit our results to the GLUECoS benchmark to see how they compare to other models!

3 Experiments

3.1 Dataset

We conduct our experiments on the LinCE CALCS English-Spanish Dataset (Aguilar et al., 2020), a widely used resource for studying English–Spanish code-switching. The corpus consists of tweets from bilingual speakers in common English/Spanish speaking areas, such as Miami and New York. Each token in the dataset is annotated using the 17 universal Part-of-Speech (POS) tags and only 3 language identification tags (eng for English, spa for Spanish, or eng&spa for tokens that appear in both languages), making it suitable for joint modeling of POS tagging and Language Identification (LID). A sample sentence from the dataset is shown in Table 1.

We preprocessed the dataset by removing non-speech artifacts (e.g., hyperlinks, emojis, and placeholders such as <unidentifiable>) and ensured that all tokens are lowercase except names. However, while pre-processing the datasets, we noticed that the LID datasets only contained the LID tags, whereas, the POS datasets had both LID and POS tags. For this reason, we chose to use the POS dataset during our experimentation to avoid differences in the amount of training data available for each task. Lastly, the test datasets had no tags, only tokens, and were used for qualitative analysis. Therefore, the metrics in Table 3 were evaluated on the POS development dataset as it had both LID and POS tags. The data is split into training, development, and test sets as shown in Table 2.

3.2 Models and Fine-Tuning Approaches

All experiments are conducted using multilingual transformer-based architectures. We tokenize the input using each model's corresponding tokenizer. For multitask fine-tuning, we configured the encoder to take each model's (XLM-R, mBERT, RemBERT) existing pre-trained configurations and slightly modify them by adding two new separate classification heads, one for POS tagging and one for LID, for task-specific output in addition to a shared transformer encoder for shared learning.

We explore the following fine-tuning strategies:

- **Single-Task Learning:** We fine-tune the model on POS and LID separately using the standard cross-entropy loss and evaluate their individual performance based on the F1 score attained on the dev set. The results obtained established a baseline performance for subsequent models on each task.

- **Sequential Learning:** The model is first fine-tuned on one task, then further fine-tuned on the second task. This approach aims to evaluate the difference in the incorporation of structural language segmentation before syntactic learning, and vice versa.

- **Un-Weighted Multitask Learning:** We use trivial weights of 1.0 with the POS and LID losses to evaluate the model's unaided performance and shared learning.

- **Loss-Weighted Multitask Learning:** Kendall et al. (2018) showed that the performance of MTL models strongly depends on the relative weighting between each task's loss. We use dynamically adjusted custom weights for POS and LID losses, prioritizing the highly uncertain task in cases of task imbalance. This is done using a loss-weighting scheme known as uncertainty-based loss weighting introduced in (Kendall et al., 2018). This approach dynamically updates the weights assigned to each task's loss using the inverse of the uncertainty at any point during the training phase with Equation 1

Token	LID	POS
pero	spa	CONJ
viste	spa	VERB
las	spa	DET
cositas	spa	NOUN
que	spa	PRON
compraron	spa	VERB
para	spa	ADP
los	spa	DET
speed	eng	NOUN
bumps	eng	NOUN
?	eng&spa	PUNCT

Table 1: Sample Sentence from the POS Training and Development Datasets

Datasets	Sentences	Tokens
POS		
Training	27,893	217,068
Development	4,298	33,345
Testing	10,720	82,656
LID		
Training	21,029	263021
Development	3,332	42,383
Testing	8,289	93,349

Table 2: Dataset Splits

$$\mathcal{L} = \sum_i \frac{1}{\sigma_i^2}\mathcal{L}_i + \log(\sigma_i) \quad (1)$$

where:

- $\mathcal{L}$ is the total loss.
- $\mathcal{L}_i$ is the loss for the i-th classification task (using Cross-Entropy Loss).
- σ_i is the learned noise parameter representing the homoscedastic (or task-dependent) uncertainty of the i-th task.

By optimizing this combined loss, the model learns to balance the different task losses by adaptively adjusting the noise parameters based on the inherent uncertainty of each task.

3.3 Evaluation Metrics

We evaluate model performance using token-level accuracy and macro F1 score for both POS and LID tasks. Accuracy is used to measure overall correctness, while macro F1 helps account for imbalanced distributions in LID and POS tag classes.

3.4 Implementation Details

All models are implemented using the Hugging Face `transformers` library and trained on an NVIDIA A100 GPU. We use the Adam optimizer with a learning rate of 5×10^{-6}, a batch size of 32 for both the training and development data sets, and a maximum sequence length of 128 (sequences longer than this were truncated and shorter ones padded). Each model is trained for up to 10 epochs with a weight decay of 0.1, a fallback dropout regularization of 0.1 for RemBERT, and the default dropout for the other models.

3.5 Baseline

As a baseline, we use the mBERT model trained jointly on POS and LID. This model, though less expressive than XLM-R or RemBERT, serves as a reference point for evaluating the benefits of large-scale multilingual pretraining and fine-tuning strategies.

4 Results

4.1 POS and LID Tagging Performance

Table 3 summarizes the performance of each fine-tuning strategy on the dev set, measured by token-level accuracy and macro F1 score for both POS tagging and Language Identification (LID). The table also contains a combined macro F1 score and accuracy which is the average of the individual POS and LID values attained while evaluating the multitask models, weighted or unweighted, on the dev set.

Across most strategies, RemBERT generally achieved the highest single-task F1 scores for both POS (96.81) and LID (93.55), followed closely by XLM-R, with mBERT performing slightly lower

but still fairly close. For all three models, single-task fine-tuning consistently outperformed joint multitask learning (both weighted and unweighted) in terms of Macro F1 scores and accuracy for the primary task. The drop in F1 score for JMT strategies compared to single-task was substantial, often around 6-8 absolute points for POS and 2-5 points for LID.

Sequential learning strategies showed mixed but interesting results. The LID then POS strategy generally maintained or slightly improved POS performance compared to single-task POS for XLM-R (96.66 F1 vs. 96.65 F1) and mBERT (96.58 F1 vs. 96.60 F1), and was competitive for RemBERT (96.66 F1 vs. 96.81 F1). Furthermore, the POS then LID strategy also showed competitive LID performance, often slightly better than single-task LID, for all models (for instance, XLM-R: 93.19 F1 vs. 93.04 F1).

Lastly, comparing loss-weighted versus unweighted JMT showed insignificantly small differences in F1 scores, but the latter revealed substantial drops in accuracy compared to single-task or sequential approaches. For all models, weighted JMT improved LID F1 scores while POS F1 scores were comparable or slightly lower. For RemBERT, unweighted JMT showed a slight edge in combined F1 and accuracy. This suggests that the weighting scheme had a modest, nuanced impact, in helping balance tasks but not overcoming the overall performance deficit compared to single-task or sequential approaches.

4.2 Analysis

The experimental results in Table 3 provide several insights into fine-tuning multilingual transformers for code-switched POS and LID tagging.

Efficiency of Sequential Learning: Sequential learning strategies (LID then POS and POS then LID) generally outperformed JMT and often matched or slightly exceeded ST performance for the second learned task. For instance, the sequential training of LID then POS yielded slightly higher POS F1 scores: for XLM-R, it was marginally higher (96.66 vs. 96.65 ST POS), and for mBERT, it was marginally lower (96.58 vs. 96.60 ST POS) than single-task POS. This supports the hypothesis that first establishing language boundaries (LID) can provide a useful platform for the subsequent, more nuanced task of POS tagging within those identified language segments. Similarly, the sequential training of POS then LID showed modest improvements in LID F1 scores and accuracies compared to single-task LID for all models. Overall, this suggests that when the model masters one aspect of CS before tackling another, it becomes more effective than attempting to learn everything simultaneously from scratch in a joint

	F1		Accuracy	
	POS	**LID**	**POS**	**LID**
mBERT				
Single				
POS	96.66		97.18	
LID		92.72		98.29
Sequential				
POS - LID	–	92.84	–	98.32
LID - POS	96.58	–	97.17	–
JMT - Unweighted				
POS & LID	88.90	90.09	90.39	97.64
Combined	89.50	–	94.02	–
JMT - Weighted				
POS & LID	88.71	90.22	90.24	97.65
Combined	89.47	–	93.95	–
XLM-R				
Single				
POS	96.65		97.25	
LID		93.04		98.37
Sequential				
POS - LID		93.19		98.39
LID - POS	96.66		97.24	
JMT - Unweighted				
POS & LID	88.82	89.57	90.36	97.52
Combined	89.20	–	93.94	–
JMT - Weighted				
POS & LID	88.88	89.74	90.37	97.58
Combined	89.31	–	93.98	–
RemBERT				
Single				
POS	96.81		97.40	
LID		93.55		98.50
Sequential				
POS - LID		93.78		98.55
LID - POS	96.66		97.28	
JMT - Unweighted				
POS & LID	88.80	88.51	90.28	97.21
Combined	88.66	–	93.75	–
JMT - Weighted				
POS & LID	88.76	88.64	90.21	97.20
Combined	88.70	–	93.71	–

Table 3: POS and LID performance (Accuracy and Macro F1) on the dev set.

manner.

Impact of Loss Weighting, Task Imbalance, and Shared Syntactical Structure: Loss-weighted JMT using uncertainty-based loss-weighting (Kendall et al., 2018), showed only marginal differences compared to unweighted JMT. While it was intended to help balance tasks, particularly by giving a slight priority to LID (which has more than 5x fewer labels than POS), its impact was not substantial enough to overcome the general performance degradation observed in unweighted JMT. This suggests that the model struggled to learn distinct feature representations for both tasks simultaneously from the mixed CS signal, perhaps due to dominance from the POS task or lack of a strong shared syntactic structure. However, both shortcomings indicate that more sophisticated dynamic weighting schemes could be explored, though these experiments were outside the scope of our initial research, and appear to be a potential for future work on this study.

Ambiguous Words: Short, frequent words that can function differently or exist in both languages (e.g., "no" - English vs. Spanish, "pan" - English verb vs. Spanish noun) are common sources of errors that negatively impact model performance, and their complete disambiguation requires external lexical resources or morphological cues.

The sequential learning approach achieves the highest performance in LID tagging, likely due to the model first learning syntactic structures before segmenting by language. On the other hand, the loss-weighted strategy yields the best LID scores, suggesting that placing emphasis on LID loss can improve the model's ability to generalize across language boundaries, even for tokens from the minority language in the dataset.

Interestingly, although the standard multitask model performs competitively, it does not match the observed performance gains seen in the other two setups, indicating that POS and LID shared learning may not fully capture the incoherence in code-switched data.

4.3 Ethical Considerations and Limitations

Developing NLP systems that can accurately parse and process code-switched text inputs is essential for achieving linguistic equity. In doing so, we help ensure that bilingual communities, representing millions of speakers globally, are not excluded from the benefits of technological advancements in AI and LLMs. This was the driving ethical force and the central motivation for our research. However, it's just as important to acknowledge the ethical considerations and limitations of this work, which should be carefully read and addressed. These include:

Dataset Specificity: We used the LinCE CALCS English-Spanish dataset (Aguilar et al., 2020), which is composed of code-switched English-Spanish tweets. These tweets were collected from active Twitter accounts located in common bilingual regions such as Miami and New York. As a result, the dataset reflects a relatively narrow social and geographic subset of code-switched speakers. As a result, our findings may not apply to other language pairs, different styles and formalities of writing (e.g., UN translations), or broader sociolinguistic settings.

Annotation Bias: The annotation process was very thorough, combining machine-generated annotations, crowd-sourced annotations via CrowdFlower, and multiple rounds of in-lab verification. Nevertheless, these annotations may still contain inconsistencies or biases. Specifically, annotator errors or lenient guidelines may have introduced noise into the dataset, potentially affecting both model performance and our evaluation metrics. Although the creators of the corpus went to great lengths to filter the annotations, some degree of error/bias likely remains, and should be considered when interpreting our results.

Model Bias: Many multilingual models, despite their broad pretraining, can inherit biases present in their training data. Consequently, these biases could have an effect on their performance on code-switched text from different demographic groups or regional dialects, even within the same language pair. For example, the English and Spanish present in the training data may not cover all of the regional dialect variations of English/Spanish equally, potentially leading to biased results or poor model performance in underrepresented communities.

Limited Statistical Analysis: Due to time constraints related to the deadline of this research paper, we were unable to confidently conduct deeper statistical analyses of our results, such as significance testing (e.g., paired t-test) across the different models. While we were able to observe some positive consistent trends, such as in sequential training (POS $\rightarrow$LID), more rigorous statistical testing would have reinforced our findings and provided greater confidence in the results.

5 Conclusion and Future Work

In this paper, we explored fine-tuning strategies for improving Part-of-Speech (POS) and Language Identification (LID) tagging in English-Spanish code-switched text, using the LinCE CALCS English-Spanish dataset (Aguilar et al., 2020). We evaluated three fine-tuning approaches: sequential training, multitask training, and loss-weighted multitask training, and evaluated their performance against a baseline single-task. After investigating three different multilingual models, our experiments found that both sequential and loss-weighted training offer consistent, albeit modest, improvements over the baseline. These results suggest that sequential training offers practical advantages that could be implemented immediately with minimal overhead, specifically for (POS $\rightarrow$LID). Furthermore, the improvements observed in loss-weighted multitask training indicate that incorporating task-specific loss balancing can meaningfully affect tagging performance, laying a solid foundation for future work. Overall, our results support the original hypothesis that task ordering and loss weighting are promising strategies for improving tagging in code-switched text.

In future work, we plan on exploring additional language pairs beyond English-Spanish, such as Hindi-English or Arabic-French, to assess the generalizability of our fine-tuning methods. Furthermore, we plan on exploring new code-switched datasets, specifically more formal texts, such as official UN translations or formal bilingual literature, to evaluate if style and text formality have an effect on code-switched tagging. Finally, we are interested in whether applying more morphological features, such as timings, gender, and mood, would increase the accuracy of POS and LID tagging.

Ultimately, we view this work as a step towards building a more equitable and linguistically inclusive future, one where NLP systems reflect the diverse ways people use language.

Acknowledgments

We would like to thank Professor Wein for her guidance throughout the course and her support on this research project. We are also grateful to Douglass Hall from the Amherst College IT Department for all his help with using the high-performance computing (HPC) cluster. Our thanks also go out to the developers and maintainers of the LinCE CALCS English-Spanish dataset. Finally, thank you to all of our peers and faculty who provided feedback on our project!

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813.

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. Association for Computational Linguistics.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23. Association for Computational Linguistics.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008. Part-of-Speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics.

Victor Soto and Julia Hirschberg. 2018. Joint part-of-speech and language ID tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10. Association for Computational Linguistics.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979. Association for Computational Linguistics.