MBART-50 Based Transfer Learning for Bidirectional MSL-Spanish Translation

Miguel Bethencourt Mayedo[1,*]

[1]Universidad de Camagüey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, 70100, Camagüey, Cuba

Abstract

This paper describes my submission to the MSLG-SPA 2026 shared task at IberLEF 2026, which focuses on bidirectional translation between Mexican Sign Language (MSL) glosses and Spanish. Due to the low-resource nature of the task (only 489 official training pairs), my approach leverages transfer learning utilizing the pre-trained multilingual sequence-to-sequence model mBART-50 Large. I implemented a four-phase training pipeline consisting of: (1) domain adaptation using the German PHOENIX-2014-T sign language dataset to establish a baseline understanding of spatial grammar; (2) large-scale synthetic data generation via rule-based back-translation of 10,000 Spanish sentences from the OPUS corpus into pseudo-glosses; (3) 5-fold cross-validation fine-tuning on the official competition dataset; and (4) inference employing optimized beam search decoding strategies. The final official evaluation yielded the following results: my system achieved a COMET score of 0.6813 and a BLEU score of 16.6254 in the MSLG-to-Spanish translation subtask (Subtask A), and a BLEU score of 11.7564 in the Spanish-to-MSLG translation subtask (Subtask B). These outcomes highlight both the potential of pre-trained models for sign language translation and the methodological limitations of my current approach. Specifically, my reliance on a standard full fine-tuning process resulted in suboptimal parameter adaptation. I analyze these limitations and propose the integration of Parameter-Efficient Fine-Tuning (PEFT) techniques, particularly Low-Rank Adaptation (LoRA), to mitigate overfitting risks and improve generalization in future low-resource sign language translation tasks.

Keywords

Mexican Sign Language, Machine Translation, mBART-50, Transfer Learning, Low-Resource NLP, Data Augmentation

1. Introduction

The MSLG-SPA 2026 shared task establishes a benchmark for bidirectional translation between Mexican Sign Language Glosses (MSLG) and Spanish (SPA) [1, 2]. The development of translation systems for sign languages represents a step towards technological inclusion for Deaf communities [3]. However, unlike traditional spoken language translation, Sign Language Translation (SLT) introduces unique morphological and syntactic complexities. Sign languages are visual-gestural languages that rely heavily on spatial grammar, non-manual markers (such as facial expressions and body posture), and simultaneous articulation, which lack direct equivalents in linear, spoken languages like Spanish [4].

Glosses serve as a written, sequential approximation of sign language, capturing the core lexical meaning but often omitting spatial and simultaneous information [4]. Consequently, the task presents two distinct computational challenges: translating lossy, grammatically divergent MSL glosses into fluent Spanish text (Subtask A: MSLG2SPA), and projecting linear Spanish text into the highly compact, topic-prominent structure of MSL glosses (Subtask B: SPA2MSLG).

The primary challenge in this shared task is data scarcity. The official training dataset consists of only 489 aligned sentence pairs, a quantity insufficient for training modern deep learning architectures from scratch. To address this, I framed the problem within a transfer learning paradigm, relying on a pre-trained multilingual model (mBART-50) and applying zero-shot cross-lingual capabilities and synthetic data augmentation techniques to mitigate the risk of overfitting.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

bethencourt.may@gmail.com (M. B. Mayedo)

0000-0001-8873-6802 (M. B. Mayedo)

My approach consists of two main components:

1. I propose a four-phase transfer learning pipeline that bridges the linguistic gap between German sign language datasets and Mexican Sign Language.
2. I use a rule-based pseudo-gloss generator that simulates MSL grammatical structures from standard Spanish text to artificially expand the training corpus.

2. Related Work

The field of Machine Translation (MT) has evolved, transitioning from Statistical Machine Translation (SMT) [5] to Neural Machine Translation (NMT), which is the current state-of-the-art [6]. NMT models, particularly those based on the Transformer architecture, have shown success in modeling long-range dependencies and generating fluent text [7].

However, the application of NMT to Sign Language Translation (SLT) remains a developing field characterized by data sparsity. Early SLT systems relied on small, domain-specific corpora, such as the PHOENIX-2014-T dataset, which consists of German weather forecasts translated into Deutsche Gebärdensprache (DGS) glosses [8]. While these datasets established a baseline for SLT, models trained exclusively on them struggle to generalize outside their narrow domains.

To combat the low-resource nature of SLT, recent approaches have increasingly adopted transfer learning and multilingual pre-training. For instance, recent studies have shown the effectiveness of fine-tuning large pre-trained language models for bidirectional SLT [9], as well as scaling pre-training to achieve zero-shot capabilities [10]. Models like mBART (Multilingual Bidirectional and Auto-Regressive Transformers) and mT5 [11] have demonstrated that language models pre-trained on diverse monolingual corpora can learn latent representations. By initializing SLT models with these pre-trained weights, researchers can effectively transfer syntactic and semantic knowledge to low-resource domains. My work builds upon this paradigm, utilizing mBART-50 as the foundational architecture to leverage its extensive pre-training on 50 spoken languages [12].

3. Data and Linguistic Characteristics

Understanding the structural disparities between MSL glosses and Spanish is important for designing an effective translation model. The official dataset provided by IberLEF 2026 consists of 489 aligned sentence pairs. This section details the linguistic features of the dataset and my data augmentation strategies.

3.1. Linguistic Disparities

The grammatical structure of MSL differs from Spanish. Key differences observed in the dataset include:

- **Word Order:** Spanish typically follows an SVO (Subject-Verb-Object) structure, whereas MSL often employs a topic-comment structure or OSV (Object-Subject-Verb) ordering to establish spatial reference before describing an action.
- **Omission of Functional Words:** MSL glosses often omit articles (el, la, los), prepositions, and copular verbs (ser/estar) that are mandatory for fluent Spanish syntax.
- **Verb Inflection:** While Spanish verbs are conjugated for tense, mood, and person, MSL verbs are often uninflected in their glossed form, relying instead on temporal markers or context to convey tense.

Furthermore, the dataset retains specific annotation conventions that the model must learn to process, including hyphens for multi-word signs (e.g., `YA-VEO`), plus signs for compound signs (e.g., `MAMÁ+PAPÁ`), number signs for loan words, and the `dm-` prefix for fingerspelling.

3.2. Synthetic Data Generation

Given that 489 pairs are insufficient to fine-tune a large neural network, I engineered a synthetic data augmentation pipeline. I sampled 10,000 Spanish sentences from the open-source `opus_books` corpus [13]. To convert these sentences into pseudo-MSL glosses, I developed a deterministic, rule-based script (`glosado.py`) utilizing the `spaCy` NLP library [14].

The script simulates the lossy nature of sign language glossing by applying the following transformations:

1. **Lemmatization:** Reducing all verbs to their infinitive forms and nouns to their singular roots.
2. **Stopword Removal:** Stripping the text of articles, prepositions, conjunctions, and auxiliary verbs.
3. **Formatting:** Removing all punctuation and converting the resulting token sequence to uppercase to mimic standard gloss notation.

While these pseudo-glosses do not capture the complex spatial reordering of true MSL, they effectively simulate the "denoising" challenge the model must overcome during translation, serving as an intermediate training step.

4. Methodology and Architecture

4.1. Model Selection: mBART-50

I selected mBART-50 Large as the backbone architecture for my system. Developed by Facebook AI Research, mBART-50 is a sequence-to-sequence model utilizing a standard Transformer encoder-decoder architecture. It was pre-trained by denoising monolingual corpora across 50 languages using span masking and sentence permutation objectives.

With approximately 600 million parameters, mBART-50 possesses a structured latent space. I hypothesized that its understanding of Spanish morphology (acquired during pre-training) would enable it to effectively "reconstruct" fluent Spanish sentences from the compressed, grammatically sparse MSL glosses.

4.2. Four-Phase Training Pipeline

To improve the transferability of pre-trained weights to the MSLG-SPA task, I implemented a four-phase training curriculum. Figure 1 illustrates this architectural workflow.

Phase 1: Domain Adaptation (PHOENIX-14T). I hypothesized that sign languages, despite regional differences, share universal typological features regarding information density and grammatical omission. I conducted a structural warm-up by fine-tuning the base mBART-50 model on approximately 8,000 pairs from the German PHOENIX-2014-T dataset (translating DGS Glosses to German text).

Phase 2: Denoising Autoencoder Training. Using the 10,000 synthetically generated pseudo-gloss pairs, I trained the model to project lossy representations back into fluent Spanish. This phase exposed the model to the specific vocabulary of Mexican Spanish and the syntactic constraints of the shared task.

Phase 3: Target Fine-Tuning. The model was subsequently fine-tuned exclusively on the 489 official MSLG-SPA pairs provided by the organizers. To validate performance internally, I employed a 5-Fold Cross-Validation strategy.

Phase 4: Optimized Inference. Inference was executed using beam search decoding. A beam width of 5 (`num_beams=5`) was selected to explore a wider hypothesis space while avoiding over-penalization of shorter sequences. Early stopping was enabled to prevent the generation of repetitive tokens.

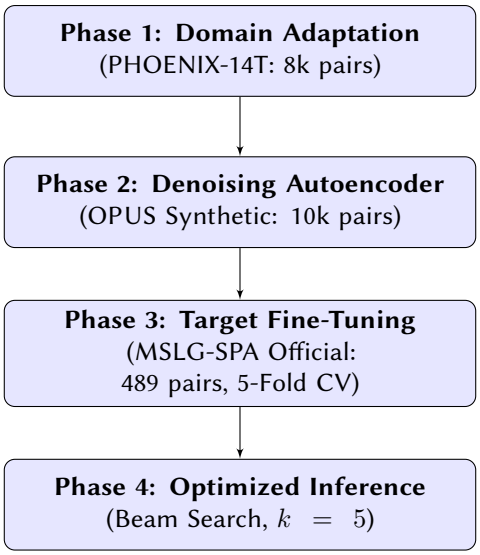

Figure 1: The proposed four-phase transfer learning pipeline utilizing mBART-50.

4.3. Hardware Constraints and Optimization

Training a 600-million parameter model requires computational resources. Due to hardware constraints, direct fine-tuning was impossible without encountering Out-Of-Memory (OOM) errors. I implemented several optimization techniques to bypass these limitations:

- **Mixed Precision (FP16):** I utilized 16-bit floating-point precision, effectively halving the memory footprint of the model activations and weights while accelerating computation on Tensor Cores.
- **Gradient Checkpointing:** Rather than storing all intermediate activations during the forward pass, I recomputed them dynamically during the backward pass.
- **Gradient Accumulation:** I restricted the physical batch size to 2 and accumulated gradients over 4 consecutive steps before executing the optimizer update, mathematically simulating an effective batch size of 8.

5. Results and Theoretical Limitations

5.1. Official Evaluation Results

My system's performance was evaluated against the official blind test set provided by the organizers. For Subtask A (MSLG2SPA), my system achieved an official COMET score of 0.6813, alongside a BLEU score [15] of 16.6254 and a chrF score [16] of 45.7760. The overall Task Score was -0.5103. For Subtask B (SPA2MSLG), the system achieved an overall Task Score of -0.0644, alongside a BLEU score of 11.7564, a METEOR score [17] of 0.3751, and a chrF score of 53.4845.

The COMET (Crosslingual Optimized Metric for Evaluation of Translation) metric [18] leverages pre-trained neural models to predict human judgments of translation quality, focusing on semantic equivalence rather than exact n-gram overlap. An official score of 0.6813 indicates a strong degree of semantic preservation and fluency, suggesting that the model successfully leveraged the synthetic pre-training phase to reconstruct grammatically correct Spanish from the sparse glosses.

5.2. The Overfitting Paradigm

Despite the positive semantic metrics (COMET), my methodological approach revealed severe theoretical limitations inherent in standard full fine-tuning. Updating all 600 million parameters of mBART-50 on a dataset of only 489 pairs is highly inefficient and prone to instability.

This imbalance between parameter count and dataset size leads to two phenomena:

1. **Pre-trained Knowledge Loss:** The large gradients generated by the small, highly specific target dataset aggressively overwrite the multilingual knowledge acquired during pre-training, causing the model to lose its foundational understanding of Spanish syntax.
2. **Overfitting:** The model rapidly memorizes the specific vocabulary and exact phrasing of the 489 training pairs. Consequently, it loses the ability to generalize to unseen test instances or natural variations in MSL glossing.

My observations during Phase 3 training confirmed rapid validation loss degradation after only a few epochs. Standard full fine-tuning simply lacks the necessary regularization mechanisms to map a highly divergent grammar (MSLG) into a high-resource language space (Spanish) when bounded by extreme data scarcity. This explains why the n-gram overlap metrics (like BLEU) remained average, even while the semantic intent (COMET) was reasonably preserved.

6. Reproducibility

To adhere to open science standards and ensure the reproducibility of these results, the source code, including the rule-based augmentation script (`glosado.py`) and the training notebooks, has been made publicly available via GitHub: https://github.com/mikebeth/MSLG-SPA-2026.

Furthermore, the foundational models and datasets utilized in my pipeline are publicly accessible: the base `mBART-50 Large` architecture can be acquired via the HuggingFace Model Hub, the synthetic data was derived from the OPUS corpus [13], and the domain adaptation was conducted using the open-source PHOENIX-2014-T dataset [8].

7. Conclusion and Future Work: The Case for LoRA

In this paper, I presented a pipeline for bidirectional translation between Mexican Sign Language glosses and Spanish, leveraging the mBART-50 architecture and synthetic data augmentation. While my system established a baseline performance highlighted by a highly competitive COMET score of 0.6813—which proves the model's capacity to preserve core semantic meaning—the core limitation lies in the parameter update strategy under data scarcity.

Future iterations of SLT systems operating in low-resource environments must abandon full fine-tuning in favor of Parameter-Efficient Fine-Tuning (PEFT) methodologies. Specifically, the integration of **Low-Rank Adaptation (LoRA)** [19] represents a viable path forward. LoRA operates by freezing the pre-trained model weights and injecting trainable rank decomposition matrices into the Transformer's self-attention layers.

By drastically reducing the number of trainable parameters, LoRA regularizes the training process. This would allow the model to learn the unique morphological mapping between MSL and Spanish without destabilizing the foundational linguistic representations embedded within mBART-50, ultimately resolving the overfitting paradigm and enabling improved structural generalization in future MSLG-SPA translation efforts.

Declaration on Generative AI

Generative AI services such as Chat DeepSeek were employed to refactor, optimize, and improve the codebase. Following the use of these services, all generated content was reviewed and edited as necessary. Full responsibility for the publication's content is assumed.

References

[1] A. Y. Rodríguez-González, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, M. G. Sánchez-Cervantes, Á. Díaz-Pacheco, R. Aranda, S. Fajardo-Flores, C. H. Martínez-Rodríguez, Overview of mslg-spa

at iberlef 2026: Bidirectional translation between mexican sign language glosses and spanish, Procesamiento del Lenguaje Natural 77 (2026).
[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.
[3] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, et al., Sign language recognition, generation, and translation: An interdisciplinary perspective, in: The 21st International ACM SIGACCESS Conference on Computers and Accessibility, 2019, pp. 16–31.
[4] K. Yin, A. Read, A. Báez, E. Kamata, et al., Including sign languages in natural language processing, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 7346–7360.
[5] P. Koehn, F. J. Och, D. Marcu, Statistical phrase-based translation, in: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003, pp. 48–54.
[6] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[8] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7784–7793.
[9] M. De Coster, M. Van Herreweghe, J. Dambre, Finetuning pre-trained language models for bidirectional sign language translation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023.
[10] B. Zhang, et al., Scaling sign language translation, in: Advances in Neural Information Processing Systems, 2024.
[11] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498.
[12] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, arXiv preprint arXiv:2008.00401 (2020).
[13] J. Tiedemann, Parallel data, tools and interfaces in opus, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 2214–2218.
[14] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear 7 (2017) 411–420.
[15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[16] M. Popović, chrf: character n-gram f-score for automatic mt evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, pp. 392–395.
[17] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
[18] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, Comet: A neural framework for mt evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2685–2702.
[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).