Enhancing Low-Resource Machine Translation via Inter-linear Text Pretraining and Hybrid Random Embeddings

David Pineda
Brigham Young University
Machine Translation Lab
pined1@byu.edu

Abstract

Lower resource machine translation remains an overlooked challenge, due to the scarcity of bilingual training data. This paper introduces a new training method to help computers translate rare languages, such as kekchi, by using word-by-word guides and a mix of smart and random word clues. We use the IBM Model 2 to extract word-level alignments to construct Interlinear Glossed Text (IGT) triples, injecting explicit linguistic bias during the pretraining process. In order to further enhance rare word translation, we introduce a hybrid scheme in the embedding layer, combining pretrained frequent token vectors with random rare token initialization. Our evaluations on a Spanish-Kekchi corpus shows significant improvements in BLEU, COMET, CHrF++ scores over our baseline models. These results promise linguistically informed pretraining for lower resource neural machine translations.

1 Introduction

Currently, neural machine translation (NMT) for lower resource languages, is a major challenge due to the scarcity of parallel corpora. Kekchi (Q'eqchi'), an indigenous Mayan language spoken in Guatemala, lack massive bilingual datasets, that are required for current neural machine translation systems. In this study, we focus on translating from Spanish (source) to Kekchi (target), an indigenous Mayan language. As a result, models suffer from poor word alignments, vocabulary sparsity, and a limited grammatical generalization (Tokarchuk and Niculae, 2024).

There have been recent advances in multilingual pretraining and large language models improving translation for medium and high resource languages. These techniques fall short in their ability to generalize a low resource scenario due to their reliance on high computational resources (Shliazhko et al., 2024; Ghosal et al., 2023)

In this paper, we propose and demonstrate a twofold approach:

- **Linguistically Informed Pretraining:** We construct Interlinear Glossed Text (IGT) triples using a IBM Model 2 alignment, which injects explicit syntactic and morphological structure into the pretraining stage. A gloss provides a morpheme-by-morpheme, or word-by-word, annotation that explains the grammatical structure of a sentence, often serving as an intermediate linguistic representation.
- **Rare Word Embedding Enhancement:** We introduce a hybrid random embedding scheme, which will initialize rare tokens with randomized vectors, while retaining their pretrained frequent tokens embeddings.

Our contributions are the following:

- We develop a gloss based strategy in pretraining. This improves structural and grammatical modeling for low resource translation.
- We propose a hybrid embedding approach for rare word translation challenges.
- We evaluate our novel method on Spanish - Kekchi corpus, demonstrating improvements across multiple evaluation metrics.

2 Related Work

Multilingual Pretraining. Improvements have been demonstrated in medium and high resource language pairs with massive multilingual models like NLLB (Team et al., 2022) and mGPT (Shliazhko et al., 2024). However, their performance still remains limited for low resource settings.

Rare Word Translation. Rare word translation remains a major challenge in neural machine translation (Khandelwal et al., 2021) proposed a k-Nearest Neighbor Machine Translation (k-NN MT)

technique that augments translation models with a data store of previously seen context-target pairs. At inference, it retrieves similar translation contexts using cosine similarity, improving accuracy particularly for rare tokens. While highly effective, this approach requires large external memory and slow retrieval operations, making it less suitable for low-resource or real-time deployments.

Hybrid Random Embeddings. Tokarchuk and Niculae (Tokarchuk and Niculae, 2024) found that randomly initialized embeddings, especially for rare words, can sometimes outperform pretrained ones in continuous-output neural machine translation. To leverage the strengths of both approaches, they introduced a hybrid strategy that blends pretrained and random embeddings using a weighted combination. This method improves representation quality for infrequent tokens while preserving the semantic structure learned from frequent ones.

Linguistically Informed Pretraining. Zhou et al. (Zhou et al., 2020) demonstrated that encoding grammatical and morphological features via glossed sequences improves translation quality in low-resource settings. These glosses abstract away from surface forms and provide a structured representation that highlights tense, aspect, and other features often lost in direct token-level translation.

Our Approach. Informed by these approaches, we construct gloss triples that associate a source language word with both a gloss (a linguistically motivated alignment label derived from Kekchi) and a Kekchi translation equivalent. For example:

- `yo` → `in` → `in`
- `quiero` → `k'ut` → `k'ut`
- `comer` → `wa'` → `wa'`

From these word-level alignments, we create training triples such as:

- (`yo`, `in`, `in`)
- (`quiero`, `k'ut`, `k'ut`)
- (`comer`, `wa'`, `wa'`)

Explanation.
`in` = 1st person pronoun in Kekchi
`k'ut` = want (infinitive/finite depending on context)
`wa'` = to eat
These terms would appear in actual Kekchi sentences like:

- Spanish: *yo quiero comer*
- Kekchi: *in k'ut wa'*

These triples combine syntactic abstraction, semantic meaning, and cross-lingual grounding. We generate them using IBM Model 2 alignments from parallel Spanish–Kekchi data and use them as input for pretraining in a format resembling Interlinear Glossed Text (IGT). We then integrate this gloss-based pretraining with hybrid rare word embeddings, aiming to enhance performance on low-resource translation tasks by improving representation of rare or morphologically rich tokens.

3 Methodology

Our goal described in this paper is to enhance low resource machine translation by integrating structured linguistic information and improve rare word representation in the embedding layer. Our approach presents a two stage training approach: first, we fine tune a mBART-50 model on Spanish and Kekchi parallel corpus; secondly, we introduce pretraining based on their Interlinear Glossed Text (IGT) triples derived from statistical alignments. Additionally, we apply a hybrid method to embed strategy which combines pretrained vectors with randomized initializations for rare words.

Here, we describe the three key components of our proposed training strategy for improving low-resource machine translation. Each component contributes to enhancing performance in different ways:

1. **Baseline Fine-Tuning:** We fine-tune an mBART-50 multilingual transformer model on Spanish–Kekchi sentence pairs to establish a performance baseline.

2. **IGT Pretraining:** We introduce linguistic structure into the model by pretraining on interlinear glossed text (IGT) triples generated using IBM Model 2 statistical alignments.

3. **Hybrid Embedding Initialization:** We improve rare word translation by initializing the model's embeddings with a hybrid of pretrained vectors and random Gaussian noise.

Each component of this training framework is described in detail below.

3.1 Dataset Overview

We conduct our experiments using a Spanish–Kekchi parallel corpus provided by the BYU Machine Translation Lab. The dataset consists of 164,903 aligned sentence pairs translated from religious texts (LDS domain), and serves as the foundation for both the baseline fine-tuning and gloss-pretraining experiments.

3.2 Baseline Fine-Tuning

We first fine-tune an mBART-50 multilingual model directly on the Spanish–Kekchi parallel corpus. The model is trained to translate standard aligned sentence pairs without any additional structural supervision, serving as the baseline system.

3.3 IGT Pretraining via IBM Model 2 Alignment

To inject structured linguistic knowledge into our translation model, we pretrain it using Interlinear Glossed Text (IGT) triples constructed from IBM Model 2 alignments. These alignments account for both lexical correspondences and differences in word order between Spanish source sentences and Kekchi translations.

Each gloss triple includes a source word, an inferred gloss term, and the corresponding target word. For example:

- (`Dios`, `xk'ut`, `Laq Dios`)
- (`creó`, `xk'ut`, `xk'ut`)
- (`el`, `la`, `la`)
- (`mundo`, `utzil`, `utzil`)

During pretraining, we construct input sequences by concatenating the Spanish source sentence with its corresponding Kekchi gloss line, formatted as:

```
[SRC] Dios creó el mundo. [GLOSS] xk'ut
                la utzil.
```

The model is then trained to predict the original Kekchi translation. This structured input encourages the model to learn morphological and syntactic correspondences between Spanish and Kekchi before fine-tuning.

Following IGT pretraining, we fine-tune the model on the original Spanish–Kekchi parallel corpus using the same data, training time, and hyperparameters as the baseline model. The IGT dataset and the raw dataset are thus equivalent in both size and content.

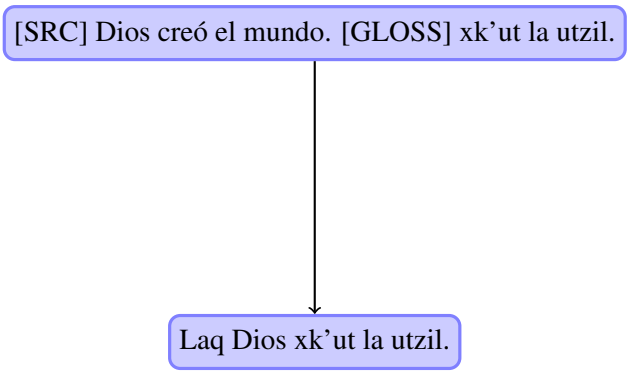

Figure 1: Input format using gloss triples to translate from Spanish to Kekchi. The gloss line represents word-aligned Kekchi words derived from statistical alignment, and the target is a fluent Kekchi translation.

3.4 Hybrid Random Embeddings

To address a vocabulary sparsity in our dataset, we introduce a hybrid word embedding initialization strategy. We establish that frequent words retain pretrained embeddings, while rare words are then assigned random Gaussian vectors.

Let V denote the vocabulary in our set, and $f(w)$ the frequency of token $w \in V$. For a threshold τ (e.g., $\tau = 100$), the embeddings are defined as:

$$\mathbf{e}_w = \begin{cases} \mathbf{e}_w^{\text{pretrained}} & \text{if } f(w) \geq \tau \\ \mathcal{N}(0, \sigma^2 \mathbf{I}) & \text{if } f(w) < \tau \end{cases}$$

where:

- $\mathbf{e}_w^{\text{pretrained}}$ is the pretrained embedding (e.g., from mBART-50),
- $\mathcal{N}(0, \sigma^2 \mathbf{I})$ is a Gaussian random vector with standard deviation $\sigma = 0.02$,
- d is the embedding dimension (e.g., $d = 768$).

Our final embedding matrix $\mathbf{E}$ combines both frequent and rare word embeddings:

$$\mathbf{E} = \mathbf{E}_{\text{freq}} \cup \mathbf{E}_{\text{rare}}$$

Example. Consider the following tokens:

Token	Frequency	Embedding Type
`"dog"`	150	Pretrained
`"run"`	180	Pretrained
`"k'ux"`	3	Random (Rare word)
`"achik"`	1	Random (Rare word)

Here, `"dog"` and `"run"` are assigned pretrained embeddings, while rare Kekchi words like `"k'ux"` (heart) and `"achik"` (dream) are initialized with

random Gaussian vectors. This allows for an even distribution of rare words across the embedding space, mitigating "hubness" issues and improving rare word decoding, as shown in prior work (Tokarchuk and Niculae, 2024).

4 Experiments

This section will outline the experimental setup, evaluation metrics used, and preliminary results comparing the baseline, mBART-50, model with our gloss pretrained version on our Spanish - Kekchi low resource translation task.

4.1 Experimental Setup

All of our models were trained on Google Colab Pro using just a single NVIDIA A100 GPU. Our hyperparameters of this experiment include: Adam, a batch size of 16, learning rate of 3×10^{-5},and enabled mixed-precision training. Each of our models trained 2 epochs with early stopping based on validation loss to prevent overfitting in our training. Pretraining on gloss triples was performed for 2 epochs, matching the number of epochs used during baseline fine-tuning

The Spanish–Kekchi parallel data was then randomly split into 80% training, 10% validation, and 10% test sets.

4.2 Experimental Configurations

To isolate the impact of each proposed enhancement, we structured our experiments into four configurations. In all cases, the model is fine-tuned on the original Spanish–Kekchi sentence pairs (i.e., raw data), using the same data size, hyperparameters, and training duration.

- **Baseline Fine-Tuning:** Fine-tune mBART-50 directly on the Spanish–Kekchi parallel corpus without any pretraining.
- **+IGT Pretraining:** Pretrain mBART-50 on gloss triples formatted as `[SRC] sentence [GLOSS] glosses`, then fine-tune on the same raw parallel corpus.
- **+Hybrid Embedding (Freq. < 100):** Following IGT pretraining, replace the embeddings of rare tokens—defined as those appearing fewer than 100 times in the training corpus—with randomly initialized Gaussian vectors. All other tokens retain their pretrained embeddings.
- **+Hybrid Embedding (Freq. < 70):** As above, but lowering the rarity threshold to tokens occurring fewer than 70 times, to evaluate the sensitivity of the hybrid embedding method to the definition of "rare."

4.3 Implementation Details

These experiments were implemented using the Hugging Face Transformers library. For mixed precision training, training was performed with the `transformers` Trainer API. All of our training runs were executed on Google Colab Pro with automatic mixed precision (AMP) enabled.

4.4 Evaluation Metrics

We evaluate translation quality using five standard metrics, each capturing different aspects of output fidelity. All metrics are computed using publicly available tools with standardized configurations to ensure reproducibility.

BLEU. BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) measures n-gram overlap between system output and reference translations, incorporating a brevity penalty to discourage overly short outputs. We report SacreBLEU scores using the `13a` tokenizer and default settings.

ChrF++. ChrF++ (Popović, 2017) computes F-scores over both character- and word-level n-grams, making it particularly effective for morphologically rich languages like Kekchi. We use the implementation from the Hugging Face `evaluate` library with $\beta = 2$ to weight recall more heavily.

ROUGE-L. ROUGE-L (Lin, 2004) measures the longest common subsequence (LCS) between the generated and reference texts, capturing sentence-level fluency and structure. We report ROUGE-L F1 using the Hugging Face `evaluate` implementation.

Exact Match (EM). Exact Match calculates the percentage of outputs that exactly match the reference translation at the sentence level, using strict string equality. This metric is useful for measuring literal correctness but penalizes even minor surface variation.

COMET. COMET (Rei et al., 2020) is a learned metric that uses contextual embeddings from multilingual encoders to predict translation quality. We report COMET scores using the `Unbabel/wmt22-comet-da` model checkpoint.

5 Results

Model	BLEU	ChrF++	ROUGE-L	EM (%)	COMET
Baseline	25.43	46.95	0.41	0.50	0.6366
+IGT Pretrain	35.31	60.67	0.59	2.70	0.7304
+IGT + Hybrid (Threshold 0.9)	**38.77**	**62.79**	**0.61**	2.10	**0.7383**
+IGT + Hybrid (Threshold 0.7)	0.24	5.46	0.05	0.00	0.2238

Table 1: Evaluation results across experimental configurations. Higher scores indicate better performance.

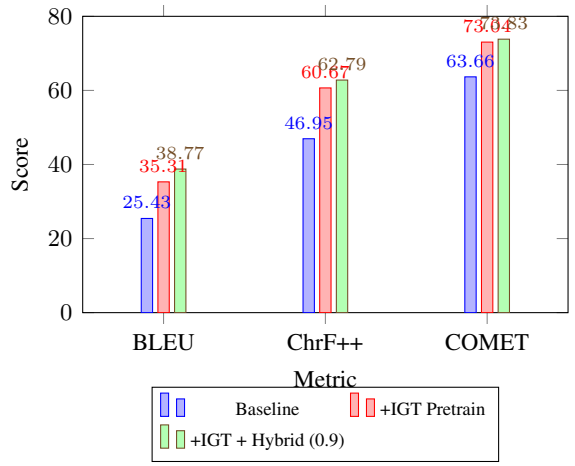

Figure 2: Comparison of BLEU, ChrF++, and COMET scores across models. Higher scores indicate better performance.

5.1 Interpretation

The Interlinear Glossed Text (IGT) pretraining demonstrates substantial improvements across all evaluation metrics. Compared to the baseline model, the gloss-pretrained model achieves a 9.9-point gain in BLEU and a 13.7-point increase in ChrF++, indicating stronger lexical and character-level modeling capabilities.

Additionally, adding hybrid random embeddings in the embedding layer with a threshold of 0.9 further improves performance, achieving the highest scores across BLEU, ChrF++, ROUGE-L, and COMET metrics. This suggests that randomizing embeddings for rare words helps reduce overfitting on rare tokens while maintaining generalization.

Further lowering the threshold to 0.7 resulted in a collapse in the model's performance: BLEU dropped to near zero, while ChrF++ and COMET scores deteriorated significantly. These findings underscore the sensitivity of hybrid embedding strategies—over-randomizing even moderately frequent words destabilizes training and severely degrades translation quality.

Overall, these results prove that linguistically informed pretraining combined with a careful rare word handling can significantly enhance lower resource neural machine translation systems.

6 Conclusion and Implications for the Field

This work demonstrates that pretraining on Interlinear Glossed Text (IGT) triples substantially improves translation quality in lower resource settings. Our approach furthers the enhancement of both lexical and semantic modeling, with a 9.9-point BLEU improvement and a 0.1 absolute COMET gain over the baseline scores.

By exposing our model to explicit grammatical structure during the pretraining stage - we introduce linguistic inductive biases that enhance fluency and alignment. Our results support recent evidence that structured linguistic data can benefit neural models even without larger scale training data (Zhou et al., 2020) (Zhang and Duh, 2021).

These findings suggest lightweight, linguistically motivated strategies as a promising path for advancing lower resource machine translation. We encourage further exploration of structured annotations not only for translation, but also for broader multilingual and cross-lingual Natural Language Processing tasks.

7 Limitations

While our findings demonstrate improvements for low-resource machine translation, several limitations should be noted:

- **Domain Specificity:** The Spanish–Kekchi dataset used in this study is drawn entirely from the religious domain. As a result, our results may not generalize to conversational, technical, or informal domains. Broader domain evaluation remains a future challenge.

8 Future Work

This work introduces a structured gloss-based pretraining method and hybrid rare-word embeddings for low-resource machine translation. While the results are promising, several future directions remain open for exploration:

- **Cross-Domain Generalization:** Our experiments are limited to religious-domain data.

Future work should evaluate whether gloss-based pretraining and hybrid embeddings remain effective on other domains such as conversational, technical, or informal text.

- **Scaling to Other Low-Resource Languages:** While Spanish–Kekchi was the focus of this study, the techniques developed here could be extended to other low-resource language pairs, especially those with rich morphology or limited parallel data. Exploring typologically diverse language pairs would provide stronger evidence of generalizability.
- **Comparison to High-Resource Settings:** To better understand whether the observed benefits are unique to low-resource contexts, it would be valuable to apply these techniques to high-resource language pairs. This could help isolate whether gloss pretraining yields diminishing returns as training data increases.
- **Automated or Learned Gloss Generation:** Currently, glosses are derived via IBM Model 2 alignments and a manually defined gloss dictionary. An exciting direction would be exploring neural methods to automatically generate glosses or morphological abstractions in low-resource settings.
- **Deeper Evaluation of Hybrid Embeddings:** Our hybrid embedding experiments show initial promise but remain preliminary. Future work should systematically evaluate embedding replacement thresholds and extend analysis across additional language pairs and tasks to assess robustness.

References

Soumya Suvra Ghosal, Soumyabrata Pal, Koyel Mukherjee, and Dinesh Manocha. 2023. Promptrefine: Enhancing few-shot performance on low-resource indic languages with example selection from related example banks. *arXiv preprint*, abs/2412.05710.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint*, abs/2207.04672.

Evgeniia Tokarchuk and Vlad Niculae. 2024. The unreasonable effectiveness of random target embeddings for continuous-output neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 653–662, Mexico City, Mexico. Association for Computational Linguistics.

Xuan Zhang and Kevin Duh. 2021. Approaching sign language gloss translation as a low-resource machine translation task. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 60–70, Virtual. Association for Machine Translation in the Americas.

Zhong Zhou, Lori Levin, David R. Mortensen, and Alex Waibel. 2020. Using interlinear glosses as pivot in low-resource multilingual machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 3765–3775, Online. Association for Computational Linguistics.

Appendix: GitHub Repository

You can view all project notebooks, code, and experiment details here:

https://github.com/pined1/Machine_Translation_Sp-Kek/tree/main

Feel free to explore, clone, and use the resources provided to build upon this work.