Machine Translation Metrics for Indigenous Languages Using Fine-tuned Semantic Embeddings

Nathaniel Krasner[1,*], Justin Vasselli[2,*], Belu Ticona[1], Antonios Anastasopoulos[1], Chi-kiu Lo 羅致翹[3]
[*]Equal Contribution, [1]George Mason University, [2]Nara Institute of Science and Technology,
[3]National Research Council Canada
nkrasner@gmu.edu, vasselli.justin_ray.vk4@is.naist.jp, mticonao@gmu.edu, antonis@gmu.edu, chikiu.lo@nrc-cnrc.gc.ca

Abstract

This paper describes the Tekio submission to the AmericasNLP 2025 shared task on machine translation metrics for Indigenous languages. We developed two primary metric approaches leveraging multilingual semantic embeddings. First, we fine-tuned the Language-agnostic BERT Sentence Encoder (LaBSE) specifically for Guarani, Bribri, and Nahuatl, significantly enhancing semantic representation quality. Next, we integrated our fine-tuned LaBSE into the semantic similarity metric YiSi-1, exploring the effectiveness of averaging multiple layers. Additionally, we trained regression-based COMET metrics (COMET-DA) using the fine-tuned LaBSE embeddings as a semantic backbone, comparing Mean Absolute Error (MAE) and Mean Squared Error (MSE) loss functions. Our YiSi-1 metric using layer-averaged embeddings chosen by having the best performance on the development set for each individual language achieved the highest average correlation across languages among our submitted systems, and our COMET models demonstrated competitive performance for Guarani.

1 Introduction

Machine translation (MT) plays a vital role in language revitalization efforts by making Indigenous language content more accessible, preserving cultural knowledge, and supporting educational initiatives that connect younger generations with their linguistic heritage. In recent years, interest in MT for Indigenous languages has grown, particularly through the AmericasNLP Shared Task in Machine Translation, which began in 2021 (Mager et al., 2021).

Due to the time-consuming and expensive nature of human annotation, automatic evaluation metrics have become essential proxy for assessing translation systems during the development cycle. These metrics offer quick, consistent, and cost-effective evaluation compared to human assessment. However, traditional metrics, such as BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2015), were designed and developed to evaluate MT systems for written and instructional languages. The distinctive features of traditionally spoken languages—e.g. polysynthetic morphology, extensive morphological variation, and non-standardized spelling—present particular challenges for metrics that rely mainly on exact matching at lexical or character level, especially when these metrics have not been specifically trained or tested in such languages. On the other hand, language representation based metrics, such as YiSi-1 (Lo, 2019), BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020), MetricX (Juraska et al., 2023), etc, require large volume of data to train the underlying language representation, which is not available for low-resource languages, like the Indigenous languages around the world.

The AmericasNLP 2025 Shared Task on Machine Translation Metrics for Indigenous Languages directly addresses this challenge, encouraging participants to develop metrics tailored to evaluate translations from Spanish into three Indigenous languages: Guarani, Bribri, and Nahuatl. The goal of the shared task is to explore and enhance MT evaluation approaches for these underrepresented languages, building upon both traditional and newer evaluation methods.

To this end, we present our approach to the shared task, leveraging recent advancements in multilingual semantic embeddings. Our contributions include:

1. Fine-tuning the Language-agnostic BERT Sentence Encoder (LaBSE; Feng et al., 2022) specifically for Indigenous languages, enhancing its ability to semantically represent translations into Guarani, Bribri, and Nahuatl.

2. Integrating these fine-tuned LaBSE embeddings into the YiSi-1 semantic similarity met-

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 100–104
May 4, 2025 ©2025 Association for Computational Linguistics

ric (Lo, 2019), exploring the impact of using different layers of LaBSE embeddings on evaluation performance.

3. Developing regression-based COMET metrics (Rei et al., 2020) using our fine-tuned LaBSE as a semantic backbone, experimenting with Mean Absolute Error (MAE) and Mean Squared Error (MSE) loss functions during training.

Our results show that fine-tuned semantic embeddings can improve MT evaluation for Indigenous languages. Our YiSi-1 using the average of embeddings from the best performing 3 layers for each individual language achieves the highest average correlation across languages among our submitted metrics, and our COMET-based metrics demonstrate competitive performance for Guarani.

2 Background

BLEU The Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) metric measures the n-gram overlap between the hypothesis text and reference translation. While BLEU is language-agnostic and simple to compute, it operates on the word level, making it challenging to accurately evaluate agglutinative languages. The organizers of the AmericasNLP shared task on machine translation Mager et al. (2021) observed that many subwords appeared in both the hypothesis and reference sentences, yet complete words frequently did not, leading them to question the usefulness of BLEU as a metric for machine translation in indigenous languages.

ChrF++ ChrF++ is a refinement of the chrF metric (Popović, 2015) that calculates an averaged F-score using precision and recall of character n-grams. The "++" variant incorporates word n-grams to slightly reward exact word matches, improving correlation with human judgments. By combining these two types of n-grams, chrF++ captures both lexical and morphological information. As ChrF++ gives partial credit for matching sub-word fragments, it is more forgiving to morphological variations than BLEU.

YiSi YiSi (Lo, 2019, 2020) is a group of semantic MT evaluation metrics designed to handle varying resource levels. YiSi represents both the hypothesis and reference sentences (or only the source, for reference-free evaluation) in a common semantic vector space, and then computes similarity scores. The primary reference-based metric is YiSi-1, which is a monolingual semantic similarity metric between the hypothesis and the reference.

For each word in the hypothesis, YiSi-1 finds the most semantically similar word in the reference (via cosine similarity of token embeddings), and vice versa, and calculates a weighted F-score. In the WMT18 Metrics Task (Ma et al., 2018), YiSi-1 showed a strong correlation with human judgments for many language pairs outperforming BLEU, chrF and others.

COMET COMET (Rei et al., 2020) is a transformer-based framework for training MT evaluation models, using human-annotated data. COMET metrics use a large multilingual model as a backbone encoder, and a regression head to predict the quality score given the source, reference, and hypothesis sentences. At inference, COMET outputs a score indicating translation quality. Based on the type of evaluation data available for training, different variants can be developed, such as COMET-DA and COMET-MQM models when using Direct Assessments (DA) and Multidimensional Quality Metric (MQM) data, respectively.

LaBSE LaBSE (Feng et al., 2022) is a BERT model with CLS-pooling and dense layers on top to produce a sentence-level encoding. This encoder is trained with a contrastive translation-ranking task to align parallel sentences between over 100 languages. Unlike many other transformer-based text encoders, which often learn disjoint spaces for each language in their training set, LaBSE represents all languages in one shared space where a sentence in one language would receive a similar encoding to its translation in any other language.

3 Methodology

Our general approach consisted of adapting a multilingual language representation across different languages into Indigenous language data, which was used to feed and train two semantic MT metrics: YiSi-1 and COMET, respectively.

3.1 Multilingual Representation using LaBSE

We fine-tuned LaBSE using a contrastive learning process to align Indigenous language data with the Spanish representation space pre-trained in LaBSE. The goal behind this alignment was to inherit the high quality pre-trained knowledge of

	Guarani		Bribri		Nahuatl		Average	
Metric	Spr.	Prs.	Spr.	Prs.	Spr.	Prs.	Spr.	Prs.
YiSi-1+ per-lang-avg	0.6611	0.7196	**0.5622**	0.6244	0.6680	**0.6115**	**0.6304**	**0.6518**
YiSi-1+ cross-lang-avg	0.6611	0.7196	0.5569	**0.6300**	0.6132	0.5845	0.6104	0.6447
COMET-DA (MAE loss)	0.5597	0.7209	0.4892	0.6261	0.4963	0.5290	0.5151	0.6254
COMET-DA (MSE loss)	0.5605	**0.7234**	0.4909	0.6268	0.5036	0.5351	0.5183	0.6285
ChrF++	**0.6725**	0.6263	0.4517	0.3823	**0.6783**	0.5549	0.6008	0.5212
BLEU	0.4676	0.4056	0.4518	0.3456	0.3541	0.4061	0.4245	0.3857

Table 1: Spearman (Spr.) and Pearson (Prs.) correlation coefficients between metrics and human scores on the blind test set across the three languages: Guarani, Bribri and Nahuatl, followed by average correlations of the three languages. per-lang-avg stands for the embeddings obtained by averaging the best three layers per language, while cross-lang-avg consider the best three layers on average in the three languages (layers 4-6). Bold values indicate the best performance in each language-correlation combination.

LaBSE with the limited data available in these low-resource languages. The data used consist of the parallel data available for the AmericasNLP MT Shared Task, which covers 13 indigenous languages, and an additional corpus for Nahuatl (Gutierrez-Vasques, 2015). For this fine-tuning process, we only propagated the gradients for the encoding of the non-Spanish sentences, aiming to preserve as much of the shared representation space as possible. Our approach consisted of training LaBSE to align all the languages simultaneously, which worked better than the language-specific models. We also balanced the language distribution data by up-sampling the training data for Nahuatl, the language for which we had less data. In this way, we improved the performance of the metrics in Nahuatl, with a small trade-off in other languages. Since the downstream translation metrics require token-level embeddings, we extracted only the BERT model from LaBSE after the fine-tuning was completed, discarding the pooling layers. While LaBSE was pre-trained by contrastive alignment of the [CLS] token encoding between parallel sentences, we found that aligning the mean-pooled token encodings to be far more effective. This is likely because aligning only the [CLS] token does not properly update the encoding of the other tokens.

3.2 Metric Development

3.2.1 YiSi-1 + Fine-Tuned LaBSE

As YiSi-1 needs an embedding model to evaluate semantic similarity (Lo, 2020), we fed this metric using the obtained LaBSE representation described in the previous section. We evaluated the metric performance using the embeddings obtained from different layers, calculating the Spearman and Pearson correlations with the DA scores. For each language, we selected the three intermediate layers that yielded the best performance on the development set and obtained the token embeddings by averaging across the three layers. However, this language-specific approach risks overfitting to the development set, potentially not performing as well on the testing set. We, therefore, made another submission for which we decided to average the token embeddings from the three layers that performed the best on average in all the three languages.

3.2.2 COMET-DA+Fine-Tuned LaBSE

We trained COMET-DA models, using our fine-tuned LaBSE embeddings as the underlying representation. Given the limited amount of available development data, we applied 5-fold cross-validation to efficiently leverage all available annotations. In each fold, we trained COMET-DA on 80% of the development set, reserving the remaining 20% for validation. We experimented with training a COMET for each language, and combining the language data. The combination led to better results on the development set, so we submitted this variation.

We explored two different loss functions to optimize COMET-DA during fine-tuning: mean absolute error (MAE) and mean squared error (MSE).

4 Results

Table 1 presents the Spearman and Pearson correlation results for our four submitted metrics compared to baseline metrics across Guarani, Bribri, and Nahuatl translation tasks. In general, our YiSi-1 metric that utilizes average embeddings from LaBSE layers 4 to 6 performed the best on average, showing strong performance across languages and metrics.

For Guarani, our COMET-based metrics performed notably well on Pearson correlation, with the COMET variant trained using MSE loss achieving the highest Pearson correlation, and the MAE variant ranking second. The YiSi-1 variants achieved higher Spearman correlations than the COMET variants, but remained lower than the ChrF++ baseline.

For Bribri, YiSi-1 with layer averaging had the highest Spearman correlation, but the single best layer for Bribri had higher Pearson correlation.

Nahuatl was especially challenging. None of our submitted metrics surpassed the ChrF++ baseline for Spearman correlation. Layer-averaged YiSi-1 scored the highest of our systems for both Spearman and Pearson.

On average, the layer-averaged YiSi performed the best of our systems.

5 Conclusion

In this paper, we present our submission to the AmericasNLP 2025 Shared Task on Machine Translation Metrics for Indigenous Languages. Central to our approach was our fine-tuned LaBSE model, which provided effective multilingual semantic representations for Bribri, Guarani, and Nahuatl. We then integrated the LaBSE embeddings into YiSi-1 and COMET metrics.

Our key contributions include successfully fine-tuning LaBSE embeddings specifically for Indigenous languages, evaluating the effectiveness of embedding layer selection and averaging in YiSi-1, and training a custom COMET for Indigenous languages.

Future directions include training COMET with additional data to further enhance COMET's performance and investigating language-specific adjustments to better handle challenging languages like Nahuatl.

Acknowledgments

This work is partially generously supported by the US National Science Foundation under awards 2327143 and 2346334. This work was partially supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Award Number 2018631).

References

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Ximena Gutierrez-Vasques. 2015. Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 154–160, Denver, Colorado. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, Online. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.