wangkongqiang at MSLG-SPA@IberLEF 2026: Bidirectional Translation between Mexican Sign Language Glosses and Spanish Using Adaptive Language Framework for T5

Kongqiang Wang[1,*], Peng Zhang[1] and Qingli Tan[2]

[1]School of Information Science and Engineering, Yunnan University, 650500, Kunming, Yunnan, China
[2]College of Ecology and Environment, Yunnan University, 650500, Kunming, Yunnan, China.

Abstract

According to a recent report by the World Health Organization, there is 1 in every 5 people in the world suffering from a hearing disorder. The organization MSLG-SPA at IberLEF 2026 considers that bidirectional translation between Mexican Sign Language (MSL) glosses and Spanish is a key effective intervention to prevent these problems in Mexico. The tasks we participated in were Task 1: MSLG-SPA and Task 2: SPA-MSLG. The objective is to address a critical gap in the development of language technologies for Deaf communities in Mexico. We compare the performance of two different modeling approaches: fine-tuning a Seq2seq neural machine translation model and using adaptive language framework for T5, with the latter yielding better results. My final experimental result is Task Score -0,5835, BLEU 12,4159, BLEU z-score -0,8013, METEOR 0,3234, METEOR z-score -0,6609, chrF 47,3885, chrF z-score -0,3054, COMET 0,6520, COMET z-score -0,5664 for MSLG2SPA and Task Score -0,2560, BLEU 10,4065, BLEU z-score -0,2807, METEOR 0,3388, METEOR z-score -0,4288, chrF 50,1617, chrF z-score -0,0586 for SPA2MSLG. The code of our project can be found at https://github.com/WangKongQiang/IberLEF_2026_MSLG-SPA.

Keywords

Hearing Disorder, Bidirectional Translation, Neural Machine Translation, ALF-T5

1. Introduction

MSLG-SPA 2026 [1] is the first shared task dedicated to bidirectional translation between Mexican Sign Language (MSL) glosses and Spanish, organized as part of IberLEF 2026 [2]. This task addresses a critical gap in the development of language technologies for Deaf communities in Mexico, where publicly available datasets, benchmarks, and computational resources for MSL remain extremely limited. The task aims to establish a structured and reproducible evaluation framework for models that can both understand and generate MSL glosses, fostering the creation of future multimodal and inclusive translation pipelines. By bringing together researchers from natural language processing, sign language processing, and artificial intelligence, MSLG-SPA 2026 seeks to lay the foundation for a sustainable research ecosystem for Mexican Sign Language.

With this context, an interesting approach is to use Natural Language Processing (NLP) techniques to train different pre-trained model [3] used by the MSLG-SPA language dataset that can be used to identify their information and provide the necessary support. The MSLG-SPA 2026 task at IberLEF 2026 aims to promote the development of NLP solutions specifically for Spanish-speaking social media [4]. They propose two main areas of focus for bidirectional translation: MSLG-SPA (Task 1), SPA-MSLG (Task 2).

In this work, we present our proposed solution to Task 1 and Task 2 of the 2026 edition of MSLG-SPA. Participants are invited to develop and evaluate automatic translation systems (computational models) for:

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
wangkongqiang60@gmail.com (K. Wang); zpp1219@gmail.com (P. Zhang); tanqingli@stu.ynu.edu.cn (Q. Tan)
https://github.com/WangKongQiang/ (K. Wang); https://github.com/zpp1219/ (P. Zhang)
0009-0003-8736-6397 (K. Wang)

Table 1
The official dataset is provided for MSLG2SPA – MSL Gloss-to-Spanish. The dataset is divided into three disjoint subsets. This partitioning ensures that both subtasks share the same training data while being evaluated on non-overlapping test sets, preventing data leakage and allowing an independent and fair assessment of each translation direction.

Dataset Type	Number of Pairs	Percentages	Comments
Train Dataset	489	50%	Used for model training in both subtasks. Participants are expected to perform their own internal validation (e.g., cross-validation or hold-out splits) on this set for model selection and hyperparameter tuning.
Test Dataset	244	25%	Used exclusively for the final evaluation of the Gloss-to-Spanish subtask.
Test Dataset	244	25%	Used exclusively for the final evaluation of the Spanish-to-Gloss subtask.

- `MSLG-SPA` Translating MSL gloss sequences into Spanish text.
- `SPA-MSLG` Translating Spanish text into MSL gloss sequences.

The rest of the paper is organized as follows: In the next section, we analyze the dataset used for these tasks (**Section 2**). Our submission system will conduct automatic evaluation through the use of machine translation (MT) metrics, which are selected based on the characteristics of each translation direction. The evaluation protocol has been stipulated (**Section 3**). Then, we describe in detail our methodology for training and evaluating the models (**Section 4**). Finally, we discuss the results obtained (**Section 5**) and present our conclusions and future lines of work (**Section 6**).

2. Dataset Analysis

The task definition of this paper is mainly as follows: The MSLG-SPA 2026 shared task focuses on the automatic translation between Mexican Sign Language glosses (MSL-G) and Spanish in a bidirectional setting. Unlike spoken language translation [5], sign languages [6] present unique challenges due to their visual-gestural modality, rich morphology, and the lack of standardized written forms. Glosses serve as an intermediate symbolic representation that enables computational processing and evaluation.

The task is divided into two complementary subtasks:

- `Subtask A: MSL Gloss → Spanish (MSLG2SPA)` : Participants develop systems that take sequences of MSL glosses as input and generate fluent and semantically accurate Spanish sentences.
- `Subtask B: Spanish → MSL Gloss (SPA2MSLG)` : Participants develop systems that translate Spanish sentences into appropriate sequences of MSL glosses, capturing the semantic content and grammatical structure of MSL.

The dataset given for these two tasks consisted of a total of hundreds of individual corpus from 489 aligned sentence-level pairs (see Table 1), each sentence with a variable number of words. Each pair represents a semantically equivalent expression in both modalities, enabling participation in the two complementary subtasks of the shared task: Gloss-to-Spanish (MSLG2SPA) and Spanish-to-Gloss (SPA2MSLG). The grammatical structure of MSL differs significantly from that of Spanish. MSL glosses do not follow Spanish word order, verbal inflection, or agreement patterns, and they often omit grammatical markers that must be reconstructed in spoken-language translation. As a result, translation between MSL glosses and Spanish involves substantial reordering, inference, and abstraction, rather than direct word-by-word mapping.

The Objectives of this task are as follows :

- Establish the first benchmark for MSL gloss–Spanish translation.
- Provide a common evaluation protocol for comparing models in both directions.
- Encourage the development of data-driven and linguistically informed approaches to sign language processing.
- Support future research toward multimodal translation systems involving video, glosses, and spoken language.

3. Evaluation Protocol

Submitted systems will be evaluated automatically using Machine Translation (MT) metrics selected according to the characteristics of each translation direction. Evaluation and ranking are performed separately for each subtask, and systems are compared only within the same subtask.

For MSLG2SPA – Gloss-to-Spanish, systems in the MSLG2SPA subtask are evaluated using the following metrics: *BLEU*, *METEOR*, *chrF*, *COMET*. Since this subtask produces well-formed Spanish sentences, COMET is applied to capture adequacy and fluency using pretrained language models [7], complementing traditional n-gram–based metrics. For this subtask, a subtask-specific Global Score is computed for each system. Scores for each metric are first standardized using z-score normalization across all submitted systems. The Global Score is then obtained as the arithmetic mean of the standardized metric values. Systems are ranked in descending order of this Global Score.

For SPA2MSLG – Spanish-to-Gloss, Systems in the SPA2MSLG subtask are evaluated using the following metrics: *BLEU*, *METEOR*, *chrF*. COMET is not applied in this subtask, as MSL gloss sequences do not constitute a natural language and do not support fluency-based evaluation. As in the MSLG2SPA subtask, metric scores are standardized using z-score normalization across all systems submitted to SPA2MSLG. A subtask-specific Global Score is computed as the arithmetic mean of the standardized BLEU, METEOR, and chrF scores. Systems are ranked in descending order of this Global Score.

This approach avoids distortions caused by naïve rank aggregation and preserves meaningful score differences between systems.

4. Methodology

4.1. Neural Machine Translation

Implementation from scratch of a machine translation with Tensorflow using a Bidirectional LSTM encoder-decoder network with attention [8]. First, use a complete Python script to convert MSLG_SPA_train.txt into the format we need (Source_language_sentence \ t Target_language_sentence \ n), and output it as MSLG_to_SPA.txt. Secondly, by using another complete Python script, we can convert the file MSLG_SPA_train.txt into the format we need (Source_language_sentence \ t Target_language_sentence \ n) and save it as SPA_to_MSLG.txt. At this point, the data preprocessing task has been completed.

For the conversion of Mexican Sign Language Glosses to Spanish, we use a NMT-MSLG_SPA.ipynb. Input is a sentence (sequence) in Mexican Sign Language Glosses. Output is the corresponding sequence in Spanish. Encoder Decoder model with a Bidirectional GRU Encoder, Attention and GRU Decoder. Use MSLG_to_SPA.txt as the training set. The split dataset (comprising training and validation sets) is 0.85:0.15. Create Seq2seq neural network graph [9]. Set the loss function, optimizer and other useful tensors. Training of the network and testing network.

For the conversion of Spanish to Mexican Sign Language Glosses, we use a NMT-SPA_MSLG.ipynb. Input is a sentence (sequence) in Spanish. Output is the corresponding sequence in Mexican Sign Language Glosses. Encoder Decoder model with a Bidirectional GRU Encoder, Attention and GRU Decoder. Use SPA_to_MSLG.txt as the training set. The split dataset (comprising training and validation sets) is 0.85:0.15. Create Seq2seq neural network graph [10]. Set the loss function, optimizer and other useful tensors. Training of the network and testing network.

Two bidirectional translation processes, the conversion of Mexican Sign Language Glosses to Spanish and the conversion of Spanish to Mexican Sign Language Glosses. The scripts used in these two processes [11], the training dataset, and the test dataset can be found in Table 2.

4.2. Adaptive Language Framework for T5

Apart from the neural machine translation approach [12] mentioned above, we also experimented with the adaptive language framework for T5 approach of taking a language model and fine-tuning it with

Table 2
The scripts, training datasets, and test datasets used in the two bidirectional translation processes for our experiments.

Python Scripts	Train Dataset	Test Dataset	Description
NMT-MSLG_SPA.ipynb	MSLG_to_SPA.txt	MSLG2SPA_test.txt	The conversion of Mexican Sign Language Glosses to Spanish.
NMT-SPA_MSLG.ipynb	SPA_to_MSLG.txt	SPA2MSLG_test.txt	The conversion of Spanish to Mexican Sign Language Glosses.

Table 3
Hyperparameters for fine-tuning *models—google-t5—t5-small*, *models—google-t5—t5-base*, and *models—google-t5—t5-large* pre-trained models for bidirectional translation tasks.

Hyperparameters	Value
Optimizer	AdamW
Learning rate	5e-4
Max Tokens	128
Num Epochs	20
Batch Size	4

the aligned sentence-level pairs of the corresponding subtasks. The model we fine-tuned was a version from *models–google-t5–t5-small*, *models–google-t5–t5-base*, and *models–google-t5–t5-large* pre-trained models for bidirectional translation from texts between in Mexican Sign Language Glosses and Spanish. We chose these models due to the fact of having been trained previously for a task that shares similar characteristics to ours. Intermediate fine-tuning has been proven to improve the results of downstream tasks by prior literature [13]. The HuggingFace Transformers and Pytorch [14] libraries in Python were utilized for loading the model weights and implementing the training loop. The models were trained using an NVIDIA GeForce RTX 3090 24G GPU for a total of 20 epochs [15].

We used an AdamW Optimizer. However, this did not improve the results empirically as compared with simply normalizing he outputs of the predictions after inference [16]. Other hyperparameters are shown in Table 3.

A neural machine translation system framework that automatically learns and translates natural languages and constructed languages (conlangs) based on a small set of translation examples. This project uses transfer learning with pre-trained language models [17] to achieve high-quality translations with minimal training data.

In the example above, T5 learned a language with only 489 aligned sentence-level examples via transfer learning and never use augmentation techniques provided by ALF framework. The model has also shown positive results for big datasets, as can be seen below on a training task for MSLG to SPA and SPA to MSLG bidirectional translation.

5. Results

Using the two approaches mentioned in the prior methodology section, we came up with different models to solve the subtasks of Task 1 and Task 2 of MSLG-SPA. The results in this section are obtained from selecting the best-performing models after evaluating the different approaches and hyperparameters on the validation set. The final predictions were obtained from a test set of MSLG or SPA sentences from 244 subjects never observed during the training process and evaluated against the task's true translation direction sentences.

In the tables (Table 4 and Table 5) below, we report the relevant metrics obtained for these subtasks and compare them against the ones obtained from other participants models provided by the organizers of the competition.

Table 4
Official reports the evaluation results from various submitted models on a MSLG2SPA test set.

Model	Task Score	BLEU	BLEU z-score	METEOR	METEOR z-score	chrF	chrF z-score	COMET	COMET z-score
alf_t5-large	-0,6605	12,4159	-0,8450	0,3234	-0,7434	47,3885	-0,3879	0,6520	-0,6658
alf_t5-base	-1,0481	8,1116	-1,0990	0,2521	-1,1306	41,4193	-0,7247	0,5723	-1,2382
NeuralMachineTranslation	-1,9051	1,2024	-1,5066	0,1125	-1,8888	14,1809	-2,2612	0,4712	-1,9638
alf_t5-small	-2,1167	0,1467	-1,5689	0,1259	-1,8162	2,1951	-2,9373	0,4461	-2,1446

Table 5
Official reports the evaluation results from various submitted models on a SPA2MSLG test set.

Model	Task Score	BLEU	BLEU z-score	METEOR	METEOR z-score	chrF	chrF z-score
alf_t5-large	-0,3338	10,4065	-0,3114	0,3388	-0,5355	50,1617	-0,1546
alf_t5-base	-1,0590	3,3902	-1,3899	0,2687	-1,1414	44,4854	-0,6456
alf_t5-small	-2,0700	1,0079	-1,7561	0,1616	-2,0664	24,3526	-2,3873
NeuralMachineTranslation	-2,4649	2,1370	-1,5825	0,0939	-2,6512	15,4097	-3,1610

6. Conclusions

The results show that these approaches considered in this work were successful at modeling of the predictive subtasks, with at least one of our models outperforming other participants in most cases. We can make the following observations:

• The best-performing approach across Task 1 and Task 2 seems to be the one that uses the embeddings of the sentence-level pairs as input to a *models–google-t5–t5-large* model. At least this model trained with this approach reached the top ranking for task absolute ranking metrics in our experiment and outperformed the other participants absolute metrics across these tasks.

• Most notably, the neural machine translation method that uses bidirectional LSTM encoder-decoder network with attention obtained the good metrics for task across all models, outperforming the other participants approach by over 20% in the absolute metrics and reaching the better spot in the Task Score metrics for these tasks in our valid dataset (train dataset split into 15% of subject ids for validation).

Another finding we can conclude from these insights is that while our models achieve great results in the absolute ranking metrics, they do not perform as well for the metrics that assess MSLG-SPA performance. In our work, we did not model explicitly for a bidirectional translation scenario. We only added information about official sentence-level pairs without through data augmentation. This limitation means our models may not perform as well in real-world situations where we aim to bidirectional translation between in other language sentences.

Thus, it may be important to explore different training approaches to improve the performance of MSLG-SPA bidirectional translation. This might include directly employing online learning to predict and update the model as new sentences come in or incorporating an ensemble of models to make independent decisions about aligned sentence-level pairs and combining them for a final decision. Additionally, we may also look into more efficient implementations of the hybrid approach to minimize the disparity in true translation direction sentences compared to transformers models. These improvements are crucial when considering the deployment of our models in real-world situations and will be the focus of future work.

7. Acknowledgments

Thank you for MSLG-SPA@IberLEF 2026 organizing the competition and providing the dataset and other support, and thanks to students of Yunnan University individuals and groups that assisted in the research and the preparation of the work.

8. Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

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] A. M. Mármol-Romero, P. Álvarez-Ojeda, A. Moreno-Muñoz, F. M. P. del Arco, M. D. Molina-González, M.-T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2025: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 75 (2025).

[4] González-Barba, J. Ángel, Chiruzzo, Luis, Jiménez-Zafra, S. María, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[5] Álvarez-Ojeda, Pablo, Cantero-Romero, M. Victoria, Semikozova, Anastasia, Montejo-Ráez, Arturo, The precom-sm corpus: Gambling in spanish social media, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 17–28.

[6] J. Xiong, O. Lipsitz, F. Nasri, L. M. W. Lui, H. Gill, L. Phan, D. Chen-Li, M. Iacobucci, R. Ho, A. Majeed, R. S. McIntyre, Impact of covid-19 pandemic on mental health in the general population: A systematic review, in: Journal of Affective Disorders 277, 2020, p. 55–64. URL: https://www.sciencedirect.com/science/article/pii/S0165032720325891. doi:10.1016/j.jad.2020.08.001.

[7] A.G.Fandiño, J.A.Estapé, M. Pàmies, J.L.Palao, J.S.Ocampo, C.P.Carrino, C.A.Oller, C.R.Penagos, A.G.Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[8] D. L. Padial, D. Gómez, hackathon-somos-nlp-2023-roberta-base-bne-finetuned-suicide-es· hugging face (2023). URL: https://huggingface.co/hackathon-somos-nlp-2023/roberta-base-bne-finetuned-suicide-es.

[9] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, in: Technometrics, 55–67, [Taylor Francis, Ltd., American Statistical Association,American Society for Quality], 1970, p. 12. URL: https://www.jstor.org/stable/1267351. doi:10.2307/1267351.

[10] L. Breiman, Random forests, in: Machine Learning, volume 45, 2001, p. 5–32. URL: https://doi.org/10.1023/A:1010933404324. doi:10.1023/A:1010933404324.

[11] C. S. Perone, R. Silveira, T. S. Paula, Evaluation of sentence embeddings in downstream and linguistic probing tasks, 2018. URL: http://arxiv.org/abs/1806.06259.

[12] J. Friedman, Greedy function approximation: A gradient boosting machine, in: The Annals of Statistics, 2000, p. 29. doi:10.1214/aos/1013203451.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R.Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, in: Journal of Machine Learning Research, 12, 2011, p. 2825–2830.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, J. B. L. Fang, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, p. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[15] C. C. Liu, J. Pfeiffer, I. Vulić, I. Gurevych, Improving generalization of adapter-based crosslingual

transfer with scheduled unfreezing, 2023. URL: http://arxiv.org/abs/2301.05487.
[16] V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J. Wilson, S. Friedler, S. Luccioni, Codecarbon: estimate and track carbon emissions from machine learning computing, in: Cited on, 2021, p. 20.
[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: http://arxiv.org/abs/1907.11692. doi:10.48550/arXiv.1907.11692.