Automated Interlinear Glossing for Low Resource Languages*

George Kaceli
University of Windsor
Windsor, ON, Canada
kaceli@uwindsor..ca

Vimanga Umange
University of Windsor
Windsor, ON, Canada
umange@uwindsor.ca

ABSTRACT

We introduce a comprehensive pipeline for automated interlinear glossing tailored for low-resource languages. Our method employs a character-level encoder enhanced with relative positional encodings, an unsupervised morpheme segmentation module featuring adaptive thresholding and a forward–backward algorithm for structured marginalization, and a Transformer decoder that incorporates cross-attention over aggregated morpheme representations. Additionally, an embedded translation component supports the model by providing auxiliary semantic cues. We evaluate our approach on the benchmark datasets using metrics such as word-level accuracy and morpheme-level accuracy.

KEYWORDS

Interlinear Glossing, Word Level Accuracy, Morpheme Level Accuracy, Low Resource Langauges

1 EXPERIMENTAL SETUP AND EVALUATION

1.1 Dataset

Our dataset consists of several low-resource languages from the SIGMORPHON 2023 Shared Task [2], the data is publicly available for download. It has been used for a variety of applications in glossing for low resource languages [3–7]. All the languages combined contain around 57,000 training sentences and all the data has been carefully annotated by competent linguistics with most of the languages including their respective morpheme segmentation as well.

The dataset is comprised of source sentences, glosses, and translations. Optionally Track 2 data contains the true morpheme segmentation as well as the true gloss.

- `\t` marks the source sentence.
- `\g` indicates the gloss.
- `\l` denotes the translation.
- **Optionally Track 2 data includes:**
  - `\t` denotes the morpheme segmentation

Note that one of the languages Nyangbo does not contain a translation, for our purposes we will not be testing on that dataset as our method could not be applied there.

The dataset contains covered and uncovered tracks where the uncovered track contains the gold label gloss and the covered track does not. Furthermore, there is track 1 and track 2 data, track 1 data does not contain the ground truth morpheme segmentation, while track 2 data does. For our purposes we are focused on the uncovered track 1 data, since we are trying to learn what gloss labels to use as well as an unsupervised morpheme segmentation to be used as additional supervision for gloss generation.

*https://github.com/gfkaceli/COMP8730_Project

1.2 Setup

Our pipeline consists of four major components:

- **Encoder:** A Transformer-based character encoder processes one-hot encoded source sentences. It uses relative positional encodings to better capture contextual dependencies across the sequence.
- **Morpheme Segmentation:** An unsupervised morpheme segmentation module computes segmentation probabilities for each character, predicts an adaptive threshold from the encoder outputs (using max, mean, and variance), and uses a forward–backward algorithm for structured marginalization. The segmentation mask is used to aggregate morpheme-level representations.
- **Translation Encoder:** A translation encoder converts translation tokens into embeddings; the mean of these embeddings forms an auxiliary representation that is used as additional supervision for the prediction.
- **Decoder:** A Transformer decoder with cross-attention takes the aggregated memory (translation plus morpheme representations) to generate the gloss sequence token-by-token.

2 EVALUATION PROTOCOL AND GOLD STANDARD

We evaluate our model on four gold-standard test sets: Gitksan, Lezgi, Natugu and Tsez. We use the following metrics taken from the shared task itself. [2]:

- **Word-Level Glossing Accuracy (wacc):** The fraction of words in the test set for which the entire gloss is correctly predicted. Formally,

$$\text{wacc} = \frac{\text{Count(correctly glossed words)}}{\text{Count(all words)}}$$

- **Morpheme-Level Glossing Accuracy (macc):** The fraction of morphemes in the test set that are correctly glossed. When the number of predicted morphemes differs from the gold standard, extra predictions are either padded with a NULL token or truncated to match the gold standard:

$$\text{macc} = \frac{\text{Count(correctly glossed morphemes)}}{\text{Count(all morphemes)}}$$

These metrics provide a comprehensive evaluation of our models performance both at the word and morpheme level.

2.1 Gloss Prediction

To predict a gloss, the model needs to learn the morpheme boundaries and then classify each morpheme with a specific gloss token, the prediction is restricted to the vocabulary of the training data and the training data may not necessarily be representative of all

the possible glossing tokens available, according to the Leipzig Glossing Rules.

3 BASELINES

In our evaluation, we compare our method against several strong baselines from the SIGMORPHON 2023 Shared Task on Interlinear Glossing[2] and others [7]. These baselines include:

- **Baseline:** The Baseline for the shared task [2]
- **TU-CL Mor:** A system that leverages deterministic mappings with hard attention for morphological inflection [4].
- **TU-CL CTC:** A system that uses Connectionist Temporal Classification (CTC) to align and generate gloss tokens[4].
- **COATES:** An ensemble encoder-decoder method, something that can be directly related and compared with our method[1].
- **TeamSiggyMorph (TSM):** A system that combines linguistic insights with neural models to generate interlinear glosses.
- **Embedded Translations (ET):** Utilizing prior approached the authors of this paper decided to include the translation as additional supervision in order to predict the gloss. In their study [7] they do not publish the morpheme level accuracy so they are not included in Table 2.

These baselines represent a range of methodologies within deep learning, and they provide a robust benchmark against which our proposed method is evaluated.

4 RESULTS

We train on one language at a time, in order to assess performance on each language. During training, the model's performance is monitored on a validation set that is built exclusively from separate data to ensure unbiased evaluation. The evaluation of the test set is performed after training to report the final model performance. We then micro-average the metrics across all the reported languages. Here is a summary and comparison of our results.

Table 1: Word-Level Glossing Accuracy for Gitksan, Natugu, Lezgi and Tsez Datasets

Method	Git (%)	Ntu (%)	Lez (%)	Ddo (%)	Average (%)
ET	28.11	85.41	82.37	85.91	70.45
Tü-CL Mor	21.09	81.04	78.78	80.94	65.46
Tü-CL CTC	4.69	80.20	78.10	80.96	60.99
COATES	6.51	70.63	65.69	74.45	54.32
Baseline	16.93	42.01	49.66	73.41	45.50
TSM	–	41.82	22.91	52.46	39.06
Ours	**36.93**	**80.25**	**85.96**	**86.91**	**72.51**

4.1 Discussion of Results

Our experiments demonstrate that the proposed pipeline achieves competitive performance in low-resource interlinear glossing tasks. The improved word-level accuracy across the datasets indicates that our method effectively produces correct gloss sequences. In particular, our approach outperforms several strong baselines by yielding an average morpheme-level accuracy of **58.41%**.On the word level

Table 2: Comparison of Morpheme-Level Glossing Accuracy on Gitksan, Natugu, Lezgi and Tsez Datasets

Method	Git (%)	Ntu (%)	Lez (%)	Ddo (%)	Ave (%)
Tü-CL Mor	11.72	56.32	62.10	73.95	51.02
Tü-CL CTC	9.26	56.38	62.03	70.29	49.49
COATES	9.84	37.84	40.74	64.43	38.21
Baseline	8.54	18.47	41.62	51.23	29.97
TSM	–	41.82	22.91	53.19	39.31
Ours	**38.11**	**54.55**	**64.49**	**76.49**	**58.41**

we also outperform other baselines with a micro-averaged word-level accuracy of about **72.51%** across all languages.

Our increased average is largely due to the notable increase in accuracy in relation to the Gitskan language. The Gitskan data is only made up of 110 total sentences, with 31 in the training data, 42 for validation, and 37 for testing. The performance of our model indicates that our method can handle data scarcity pretty well and can generalize adequately despite these limitations. In other instances where the splits between training, validation, and testing are more in line with common practices, our model maintains competitive performance with the other baselines, improving accuracy in Lezgi on the word and morpheme level.

The results suggest that combining encoder–decoder architectures with robust segmentation and translation encoding is a promising strategy for low-resource scenarios. While the higher morpheme-level accuracy reflects effective per-morpheme performance, the stricter requirements of word-level accuracy highlight the challenges in fully matching the gold standard. Data scarcity, particularly in languages like Gitksan, contributes to lower generalization, whereas languages with larger datasets (e.g., Natugu and Lezgi) perform better.

REFERENCES

[1] Edith Coates. 2023. An Ensembled Encoder-Decoder System for Interlinear Glossed Text. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology.* Department of Mathematics, University of British Columbia, Vancouver, Canada..

[2] Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Toronto, Canada, 171–185.

[3] Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024. GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text. arXiv:2403.06399 [cs.CL] https://arxiv.org/abs/2403.06399

[4] Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-Through Gradient Estimation for Hard Attention. In *SIGMORPHON.*

[5] Taiqi He, Lindia Tjuatja, Nate Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun. In *Proceedings of the 2023 SIGMORPHON Workshop.* Association for Computational Linguistics, 112–118. https://doi.org/10.18653/v1/2023.sigmorphon-1.12

[6] Raphael Schwitter Martin Volk Lukas Fischer, Patricia Scheurer. 2022. Machine translation of 16th-century Latin letters into German: A case study on historical document translation. In *Proceedings of the Workshop on Language Technology for Historical and Ancient Languages (LT4HALA).*

[7] Changbing Yang et al. 2024. Embedded Translations for Low-resource Automated Glossing. In *SIGMORPHON.*