Literature Review: Overview of the Development of Automated Interlinear Glossing Techniques

George Kaceli
University of Windsor
Canada
kaceli@uwindsor.ca

Vimanga Umange
University of Windsor
Canada
umange@uwindsor.ca

ABSTRACT

Advancement of automated glossing and transformer models hold significant potential to streamline linguistic documentation. This literature review explores Automated Interlinear Glossing (AIG), examining the effectiveness of existing approaches. Placing special emphasis on complexities of glossing and addressing ambiguities in morphology and syntax

KEYWORDS

Morpheme Segmentation, Interlinear Glossing Text (IGT), Transformer Models, Low-Resource NLP, Online Database of IGT (ODIN)

1 HISTORY

Machine translation (MT) traces back to early linguistic theories, with Rene Descartes proposing a universal language in the 17th century. Interlinear Glossed Text (IGT), a subfield of MT, has been essential for language preservation, initially relying on manual annotations in ancient time before transitioning to computational approaches in modern history. However, glossing remains labor-intensive, making automation increasingly relevant. The formal development of MT began in the 1950s when Yehoshua Bar-Hillel organized the first International Conference on Machine Translation. MT evolved through three primary paradigms: Rule-Based Machine Translation (RBMT), Statistical-Based Machine Translation (SMT), and Neural Machine Translation (NMT).

Neural networks advanced MT with encoder-decoder architectures [1], allowing end-to-end training without manual feature engineering. Transformers further improved semantic representation, leading to the rise of multilingual Large Language Models (LLMs). Automated Interlinear Glossing (AIG), a subtask of MT, benefits from deep learning techniques, enabling gloss generation for under-documented languages and contributing to language preservation efforts.

2 HIERARCHY OF CATEGORIES

AIG has evolved significantly with the emergence of deep learning techniques. Early systems replaced manual annotation but were difficult to scale across languages. Attempts to incorporate annotated corpora and statistical models seemed fruitful, but simpler architectures failed to generalize. The advent of transformers and LLMs has enhanced contextual understanding and generalization.

2.1 Unsupervised Methods

Early unsupervised approaches leveraged morphological clustering and statistical heuristics to preprocess IGT data. [8] explored clustering techniques using document boundaries, while earlier methods [5, 10] relied on prefix overlap heuristics, which were computationally rigid. Cross-lingual projection [13] showed potential for low-resource glossing but struggled with alignment inconsistencies. Unsupervised methods have become more context-aware, integrating deep learning techniques. However, modern research focuses on gathering and training on large amounts of labeled data to further enhance glossing efficiency and accuracy.

2.2 Supervised Methods

Supervised learning significantly improved annotation efficiency and model accuracy. Active learning [9] reduced annotation workload by selecting informative examples for human labeling. Cross-lingual annotation transfer [11] enabled low-resource glossing using high-resource language alignments, refining morphological and syntactic projections. These methods emphasized human-in-the-loop learning, a concept that remains integral to modern neural glossing models. Transformers , when trained on labeled data, provided state-of-the-art performance in IGT, enabling greater contextual representation and accuracy.

2.3 Neural Networks and Transformers

Deep learning and transformer-based models revolutionized AIG by allowing end-to-end learning, contextualized embeddings, and multi-level sequence prediction. Early neural approaches relied on RNNs and Seq2Seq models, which encoded text into fixed-dimensional representations but struggled with long-range dependencies. [4] improved upon this by introducing straight-through gradient estimation for hard attention, enhancing morphological segmentation accuracy. Yang et al. [12] extended this work by integrating embedded translations with character-level decoding, demonstrating that translation supervision further improves glossing accuracy. The introduction of GlossLM [3], a pre-trained model designed specifically for IGT, which unlike BERT, incorporates interlinear glossed text during pretraining, enhancing morphological and syntactic understanding while reducing reliance on large labeled datasets.

This integrated hierarchy illustrates the progression of machine translation methodologies and their applications to interlinear glossing. Each paradigm has contributed to advancements in handling the complexities associated with these tasks.

3 SUMMARIES OF SUBCATEGORIES

Machine learning helps automate glossing by recognizing patterns in language data. Different techniques can be incorporated to aid in the generation of morpheme segmentation and glossing.

3.1 Summary of Unsupervised Methods

Hafer and Weiss [5] refined the Letter Successor Variety (LSV) algorithm by integrating entropy-based heuristics, improving segmentation accuracy through analysis of predecessor and successor letter distributions. Their approach was tested on the Brown corpus and others. Recognizing statistical segmentation limitations, Jurafsky [10] introduced distributional similarity and word clustering, enabling morphological induction without human input. The algorithm, adaptable to German, Dutch, and English, leveraged shared syntactic structures. Expanding beyond character-based heuristics, Moon et al. [8] developed a parameter-free induction model using document boundaries as linguistic signals. Their approach identified stems based on affix variations, filtered and clustered affixes by co-occurrence, and grouped stems sharing common affixes. Evaluated on English and Uspanteko datasets, their method improved segmentation accuracy. Building on document-level segmentation, [7] introduced a machine learning approach for automated gloss generation, by incorporating context-aware morphological segmentation. She used pattern-based machine learning models without explicit labeling, learning segmentation rules. Evaluated on IGT from endangered language documentation projects, her method scaled gloss generation for under-documented languages.

3.2 Summary of Supervised Methods

Most NLP research focuses on languages with extensive annotated data, while over 200 languages lack sufficient training resources [13]. Cross-lingual methods emerged to support NLP tasks in low-resource settings by leveraging high-resource languages [11]. Active Learning for IGT prioritizes informative linguistic examples for annotation, improving glossing accuracy with fewer labeled examples [9]. Encoder-decoder architectures encode source text into contextual representations and decode structured glosses [1], enhancing annotation efficiency and translation quality in low-resource settings. Ginn and Palmer[2] introduced robust generalization strategies for morpheme glossing in endangered languages, addressing overfitting through data augmentation, transfer learning, and fine-tuning. Their approach, tested on the ODIN dataset, improved model generalization to unseen languages. The SigMoreFun system [6] refined supervised glossing by integrating neural architectures with data augmentation, improving segmentation and glossing accuracy. Evaluated on the SIGMORPHON 2023 dataset, it demonstrated state-of-the-art performance in interlinear glossing. These advancements illustrate the evolution of supervised glossing, from cross-lingual annotation transfer to deep learning-based sequence models.

3.3 Summary of Transformers Methods

Zhao et al. [14] introduced an interlinear glossing model that uses aligned translations as weak supervision to enhance morphological segmentation and gloss prediction. Their model was tested on languages such as Lezgian, and Arapaho, demonstrating its effectiveness in bridging traditional glossing models with translation-based approaches. Girrbach [4] developed a hard-attention mechanism with straight-through gradient estimation (STGE) for AIG. Unlike soft-attention models, hard attention selects morpheme boundaries more precisely, leading to improved segmentation accuracy. This method was evaluated using the SIGMORPHON dataset. Yang et al. [12] built upon Girrbach's approach by integrating multilingual transformers, such as BERT and T5, with character-level decoding and translation supervision. This enhancement further refined segmentation and gloss prediction accuracy on the same dataset. Ginn et al. [3] introduced GlossLM, a specialized multilingual language model trained on a dataset of 450k IGT examples from 1.8k texts, the most extensive corpus available for IGT tasks, setting it apart from general-purpose models like BERT and GPT. GlossLM significantly improved morphological segmentation for low-resource languages.

4 POSITIONING OUR RESEARCH

Our research aligns with Girrbach's award-winning work at the SIGMORPHON 2023 Shared Task, which improved segmentation and gloss accuracy [4]. Subsequent studies further enhanced glossing by incorporating embedded translations and character-level decoding. Our research builds upon Girrbach's method by aiming to improve generalization, segmentation, and enhance glossing performance.

This review highlights different AIG methods. While significant progress has been made, challenges remain in adapting these systems to diverse and morphologically rich languages. Our work aims to address these concerns and make improvements on existing methodologies.

REFERENCES

[1] Kyunghyun Cho, Bart Van Merrinboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[2] Michael Ginn and Alexis Palmer. 2023. Robust Generalization Strategies for Morpheme Glossing in an Endangered Language Documentation Context. In *Proceedings of the 1st GenBench Workshop*. Association for Computational Linguistics, Singapore, 89–98. https://doi.org/10.18653/v1/2023.genbench-1.7

[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] M.A. Hafer and S.F. Weiss. 1974. Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval* 10 (1974), 371–385.

[6] 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

[7] Angelina McMillan-Major. 2020. Automating Gloss Generation in Interlinear Glossed Text. In *Proceedings of the Society for Computation in Linguistics 2020*. Association for Computational Linguistics, 355–366. https://aclanthology.org/2020.scil-1.42/

[8] Taesun Moon, Katrin Erk, and Jason Baldridge. 2009. Unsupervised Morphological Segmentation and Clustering with Document Boundaries. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore.

[9] Alexis Palmer and Jason Baldridge. 2009. Computational strategies for reducing annotation effort in language documentation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. 241–249.

[10] Patrick Schone and Daniel Jurafsky. 2001. Knowledge-Free Induction of Inflectional Morphologies. In *North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics, 1–9.

[11] Fei Xia and William Lewis. 2007. Multilingual Structural Projection Across Interlinear Text. In *Proceedings of HLT/NAACL 2007*. Rochester, NY.

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

[13] David Yarowsky and Grace Ngai. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora. In *Proceedings of NAACL 2001*. 200–207.

[14] Xingyuan Zhao et al. 2020. Automatic Interlinear Glossing for Under-Resourced Languages Leveraging Translations. *COLING* (2020).