Investigating NLP techniques or Automated Glossing of Classical and Other 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

Classical and low-resource languages face significant NLP challenges due to limited resources and complex morphologies. This research explores automated glossing with embedded translations to address these issues. By integrating historical linguistic data and advanced NLP techniques, this study aims to enhance the accessibility and interpretation of classical texts and underrepresented languages.

KEYWORDS

Natural Language Processing, Interlinear Glossing, Classical Languages, Low-Resource NLP

1 INTRODUCTION

Interlinear glossing is a linguistic annotation technique that provides word-by-word translations and grammatical information for a sentence. It is essential for understanding the structure and meaning of texts in classical and low-resource languages. For many low-resource languages, this is the only form of annotated data that is available for NLP work. Creation of glossed text is, however, a laborious endeavour and this project investigates methods to automate the process.

Recent advancements in natural language processing (NLP) offer opportunities to improve automatic glossing. Transformer-based models and multilingual embeddings, have shown promise in handling low-resource and morphologically rich languages by modeling contextual and grammatical dependencies effectively [1, 4] . Methods have also shown that using embedded translations can garner improvement in automated glossing results [1].

2 MOTIVATION

Automated glossing addresses critical challenges in low-resource NLP by facilitating linguistic research, providing tools for educators and students, and preserving under-documented languages like Arapaho. Classical languages, such as Ancient Greek and Latin, pose unique challenges due to their complex grammatical structures, but these same features offer opportunities for improved contextual understanding when modeled with neural networks [2]. By building on multilingual glossing efforts [4] and integrating hard-attention mechanisms [3], furthermore embedded translations in [1], can potentially be improved by incorporating more complex decoder architecture. Which can aid with more efficient and accurate gloss generation.

2.1 Motivating Example

For example, automated glossing can annotate Ancient Greek texts with morpheme segmentation and translations, providing researchers and students with accessible linguistic insights. Applying more complex decoder models has the potential to improve and supporting the study of low-resource languages especially those at risk of extinction.

3 PROBLEM DEFINITION

Let $M = s_1, s_2 ..... s_n$ be a corpus of language sentences, and $S = w_1, w_2, ... w_n$ be a sentence consisting of tokens. The task is to generate an interlinear gloss $G = g_1, g_2, ... g_n$ for S, where each g_i is a tuple of the form $g_i = \{ seg(w_i), morph(w_i), trans(w_i) \}$. The elements in the tuple are defined as follows:

seg(w_i): Morpheme segmentation of token w_i .
morph(w_i): Morphological annotation of w_i (e.g., case, tense).
trans(w_i): Context-sensitive translation w_i .

Examples:
Input Sentence (Ancient Greek): .
Morpheme Segmentation: -
Gloss Line: man.NOM.SG run.PRS.3SG
Translation (English): The man runs.
Input Sentence (Arapaho): Noh neihoowbeet3eiisin.
Morpheme Segmentation: nei-hoow-beet-3eiis-in
Gloss Line: and 1.NEG-want.to-be.in.jail
Translation(English): And I don't want to be in jail .

4 AUTHOR LIST

The authors of this proposal are as follows:

- **George Kaceli, University of Windsor**: George has a strong academic background and has previously worked on language processing projects and with transformer based models such as BERT and GPT.
- **Vimanga Umange, University of Windsor**: Vimanga has a strong academic background as well as three years of industry experience in software development.

The collaboration between George and Vimanga is based on their shared interests. Both members bring complementary skill sets, ensuring the project's success and thus will each contribute equally to the project.

REFERENCES

[1] Changbing Yang, Garrett Nicolai, Miikka Silfverberg. "Embedded Translations for Low-resource Automated Glossing." University of British Columbia.
[2] Aleksei Dorkin, Kairit Sirts. "TartuNLP @ SIGTYP 2024 Shared Task: Adapting XLM-RoBERTa for Ancient and Historical Languages." Institute of Computer Science, University of Tartu.
[3] Leander Girrbach. "Tü-CL at SIGMORPHON 2023: Straight-Through Gradient Estimation for Hard Attention." University of Tübingen.
[4] Shu Okabe, François Yvon. "Towards Multilingual Interlinear Morphological Glossing." Université Paris-Saclay CNRS, LISN; Sorbonne Université CNRS, ISIR.