University Of Calgary
Department Of Computer Science
CPSC 599.27, Winter 2023

Ancient Greek Machine Translation With OpenNMT & CLTK

Andrew M. Burton

January 31, 2023

Student ID: 30008738
Instructor: Dr. Katie Ovens
Teaching Assistant: Tejash Shrestha

Ancient Greek Machine Translation With OpenNMT & CLTK

ANDREW BURTON, University of Calgary, Canada

This project proposes the development of a tool intended to be used for the translation of Ancient Greek texts using the OpenNMT [3] framework and the Natural Language Processing library, CLTK [2]. The tool will leverage the corpora made available by CLTK to train the OpenNMT model, namely the Perseus Digital Library [1] and the thesaurus Linguae Graecae [5], thereby addressing the issue of inaccuracies present in current pretrained models. Text pre-processing will be performed utilizing the capabilities of CLTK, while the OpenNMT model will generate the translations and produce human-readable outputs. The tool will also consider various qualities of the original text, which will be assessed by comparing Part-of-Speech (POS) tags with the output text to ensure it meets a reasonable measure of confidence. This project aims to provide a solution that makes translational details easier to parse for novice translators processing ancient texts, thereby facilitating a deeper understanding of the language with context. The resultant tool will offer users a user-friendly and highly accurate means of translating these valuable texts, making the study of ancient languages more efficient, effective, and accessible.

CCS Concepts: • **Applied computing** → **Document management and text processing**; • **Computing methodologies** → **Machine translation**.

Additional Key Words and Phrases: datasets, neural networks, translation, text tagging

1 INTRODUCTION

This project suggests utilising OpenNMT and the Classical Languages Toolkit (CLTK) to create a tool for parsing Ancient Greek manuscripts. Classical languages and low resource languages like Latin and Ancient Greek are the focus of the CLTK suite of natural language processing technologies. A range of resources are offered by the toolkit, including corpora, tokenization tools, and morphological analysis tools. Before translation in this project, the Ancient Greek text will be preprocessed using CLTK. OpenNMT is a framework for neural machine translation that has proved effective in a number of NLP applications. The adaptable design of OpenNMT enables the fine-tuning of pretrained models on certain domains. The preprocessed Ancient Greek text will be translated in this project using OpenNMT. The project aims to create a novel tool that overcomes the shortcomings of existing Ancient Greek translation schemes. The programme will also give the user extra information by using pre-processing data like Part-of-Speech (POS) tags. Users will then be able to determine how confident they are with the tool's translations.

2 RELATED WORK

Given the impact of ancient Greek culture on modern society developing tools for analyzing the language helps provide insights for people trying to develop skills for processing texts. Some projects that have helped propel this endeavour forward include:

(1) **Eulexis** - Eulexis is a powerful lemmatization application that has been trained on a variety of greek corpora. It uses a combination of rule-based and machine learning-based techniques to perform morphological analysis and lemmatization, allowing for more accurate search results, than many models are able to produce.
(2) **Perseus Project** - The Perseus Project is a digital library for the study of the ancient world. It provides access to a variety of ancient Greek and Latin texts, as well as tools like Morpheus for searching, analyzing, and visualizing the data. It has been used for a range of research projects, including machine translation.
(3) **Preserving Personality through a Pivot Language** - Details a paper using NMT for Greek in order to preserve certain qualities of the text. [4]

Author's address: Andrew Burton, andrew.burton@ucalgary.ca, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada, T2N 1N4.

3 PROPOSED WORK

In this project, we aim to address the issue of making translations for ancient Greek accessible by making a tool that can provide learners and experts with detailed information about the generated translations.

3.1 Data Collection

The first step will involve creating source and target files for OpenNMT to use in its own training process. this will be done by utilizing python libraries such as `beautifulsoup4` and `requests` in order to scrape successfully translated texts from the Perseus Digital Library [1] and other sources as needed, in order to create target and source training data for OpenNMT.

3.2 Data Preprocessing

The first step in processing the Greek text for this project is to use the Classical Language Toolkit (CLTK) to perform preprocessing operations. CLTK provides several NLP functions that will make the Greek text easier to process, including converting the text to a uniform case and removing diacritics. By breaking the text down into smaller, more manageable chunks, the NLP models can better understand and process the text. Additionally, CLTK provides several tools for working with ancient Greek text, including functions for normalizing text, removing unwanted characters, and removing stop words. These tools are designed to help make the text more manageable and improve the accuracy of the NLP models. CLTK provides a range of NLP functions and tools that will assist in simplifying the preprocessing steps without losing accuracy in annotating the text with tags.

3.3 Training

Once the data has been prepared for OpenNMT, a model will be trained on the text data using an encoder-decoder architecture, where the encoder takes the source language text as input and the decoder produces the target language text as output.

3.4 Interaction

The application will present users with an interactive application where they can input text and receive the translations output along with labels giving the POS for each Greek word.

3.5 Analysis

In order to gauge both the effectiveness of the program and the caliber of the translations, this tool will be assessed using a number of criteria. Given that a significant portion of the data from Perseus includes prepared tags for cross validation, POS tag accuracy may be evaluated against known terms in the training set. It will be crucial to assess how this tool stacks up in terms of tagging versus a rival option like Eulexis. Another crucial source of assessment will come from human evaluation; this information may be gathered through University of Calgary students taking Greek language programs as well as by asking the instructors to assess the tool's usefulness.

REFERENCES

[1] Gregory R. Crane. 2023. Perseus Digital Library. Tufts University. http://perseus.tufts.edu

[2] Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 20–29. https://doi.org/10.18653/v1/2021.acl-demo.3

[3] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 67–72. https://www.aclweb.org/anthology/P17-4012

[4] Annie Lamar and America Chambers. 2018. Preserving Personality through a Pivot Language Low-Resource NMT of Ancient Languages. 1–4. https://doi.org/10.1109/URTC45901.2018.9244794

[5] Maria C. Pantelia. 2023. Thesaurus Linguae Graecae® Digital Library. University of California, Irvine. http://www.tlg.uci.edu