1. Ground Truth Generation for Multilingual Historical NLP using LLMs

Accession number: 20250518963

Authors: Gladstone, Clovis (1); Fang, Zhao (2); Stewart, Spencer Dean (3)
Author affiliation: (1) ARTFL Project, University of Chicago, Chicago, United States; (2) Department of History, University of Chicago, Chicago, United States; (3) Libraries and School of Information Studies, Purdue University, West Lafayette, United States

Source title: arXiv
Abbreviated source title: arXiv
Issue date: November 18, 2025
Publication year: 2025
Language: English
E-ISSN: 23318422
Document type: Preprint (PP)
Repository: arXiv

Abstract: Historical and low-resource NLP remains challenging due to limited annotated data and domain mismatches with modern, web-sourced corpora. This paper outlines our work in using large language models (LLMs) to create ground-truth annotations for historical French (16th–20th centuries) and Chinese (1900–1950) texts. By leveraging LLM-generated ground truth on a subset of our corpus, we were able to fine-tune spaCy to achieve significant gains on period-specific tests for part-of-speech (POS) annotations, lemmatization, and named entity recognition (NER). Our results underscore the importance of domain-specific models and demonstrate that even relatively limited amounts of synthetic data can improve NLP tools for under-resourced corpora in computational humanities research. © 2025, CC BY.

Number of references: 0
Main heading: Natural language processing systems
Controlled terms: Computational linguistics
Uncontrolled terms: Ground truth - Historical NLP - Language model - Language processing - Large language model - Multilingual NLP - Natural language processing - Natural languages
Classification code: 1101 Artificial Intelligence - 1102.1 Computer Theory, Includes Computational Logic, Automata Theory, Switching Theory, Programming Theory - 1106.2 Data Handling and Data Processing - 1106.7 Computational Linguistics
DOI: 10.48550/arXiv.2511.14688
Compendex references: YES
Preprint ID: 2511.14688v1
Preprint source website: https://arxiv.org
Preprint ID type: ARXIV
Database: Compendex
Data Provider: Engineering Village