Semantic Entity Recognition in Historical Texts: Extracting Information from Late Imperial China's Inscriptions on Material Infrastructure

Wangzhi Xi (KU Leuven)

Introduction

This research undertakes a Semantic Entity Recognition (SER) task, an extension of traditional Named Entity Recognition (NER), focusing on inscriptions on material infrastructure in late imperial China. SER, in this context, is not limited to identifying named entities such as persons, locations, and organizations. Instead, it encompasses a more comprehensive range of entities, including spatial, temporal, descriptive, quantitative, qualitative, and conceptual ones. This broadened scope is essential for extracting and interpreting multi-layered information in historical sources. Recent advancements in large pre-trained language models have highlighted their potential in various natural language processing (NLP) tasks, even with small labeled datasets.

This work forms a critical part of a larger project Regionalizing Infrastructures in Chinese History (RegInfra), which explores how large-scale infrastructures, contributed to regional and empire-wide integration and (re)formation in late imperial China (ca. 1000-1900 CE) (see De Weerdt, Xiong, and Liu 2020 report on its pilot project). The aim is to automatically extract information from our annotated data through NER methodologies. Our primary sources are stele inscriptions related to city walls, bridges, and roads recorded in *fangzhi* 方志, commonly translated as "local gazetteer". Chinese local gazetteers, a genre of writings that documenting geography, history, literature, and government records of a territorial unit, are often considered the best source for local information (Dennis 2015). Our current corpus consists of 467 annotated inscriptions, totaling 280,823 characters, encompassing 31,455 occurrences of 16 different entity types across various provinces in China. Preliminary experiments using classical Chinese BERT models have shown the potential of SER on our dataset.

Objectives and Challenges

The primary objective of this study is to develop a model for the automatic extraction of semantic entities for the RegInfra and InfraLives teams. Our corpus, sourced from the Database of Chinese Local Records (*Zhongguo fangzhi ku*), is being continually enriched by

ongoing annotations from the RegInfra team and members of our allied project the Lives and Afterlives of Imperial Material Infrastructure in Southeastern China (InfraLives) via the MARKUS platform. Our interest extends beyond traditional named entities to include details of infrastructure construction, deconstruction, renovation, and failure, encompassing a range of entities, such as time and place, actions with their durations, costs, and causes, information about actors including names, official titles, social categories, and social relations, details about infrastructure objects like names, parts, materials, quantities, and measurements, as well as references to deities.

The secondary objective involves a two-tiered classification approach, further categorizing some of these initially identified entities into more specific subgroups. For example, materials are classified into categories such as wood, stone, and earth, and social categories are classified into official, gentry, merchant, commoner, etc.

Processing ancient Chinese texts presents several challenges, including the use of archaic or obsolete characters and the absence of explicit word and sentence boundaries. Recent developments in named entity processing in classical Chinese writings using state-of-the-art pre-training plus fine-tuning methodologies have yielded promising results. However, most existing studies focus on standard named entities. Our project, with its historical research orientation, faces challenges such as managing labels of high granularity and uncovering implicit semantic relations between entities.

The SER model developed in this study will be incorporated into the MARKUS platform to facilitate markup and corpus building as well as macro-level analysis for historical researchers. Moreover, this task is a component within a broader modeling pipeline of our project, which includes a subsequent relation extraction task leveraging event markup on the same sources and the integration with knowledge graph methodologies.

Methodology

Our current dataset comprises inscriptions of city walls and bridges in Fujian, Shaanxi, Shanxi, Shandong, Sichuan, and Zhejiang provinces. We have converted HTML markup into datasets of aligned token-label pairs (each character as a token) using the BIO tagging schema (Begin, Inside, Outside). Currently, we use our dataset as a whole. As the dataset grows, we plan to refine our approach by potentially subdividing the dataset based on object types (city wall, bridge, road), time periods (dynasty), and locations (province), or by incorporating these as features in our models.

In preprocessing, we handled texts lacking boundary marks by using a fixed-length sliding window technique to segment each text into smaller, consistent-sized chunks, with overlaps at window boundaries to maintain continuity.

For model training, we initially established a BiLSTM-CRF (bidirectional long short-term memory and conditional random field) model as our baseline. We then explore pre-trained models of modern Chinese (CKIP BERT Base Chinese[1]) and classical Chinese (GuwenBERT[2], SikuBERT[3], and their derived RoBERTa models), also incorporating a CRF layer for inference. The models are evaluated with precision, recall, and F-1 score as our primary metrics. Additionally, we plan to test the generalizability of the models to similar datasets, such as texts related to city walls, roads, and bridges that do not belong to the genre of inscriptions.

Preliminary Results, Discussions, and Future Work

This research is in progress. The preprocessing of the existing dataset has been completed. Currently, we are in the early stage of model training and evaluation.

Our preliminary results, as an initial indication of the potential of our approach, show that BERT models consistently yield higher F-1 scores. Among the models, SikuRoBERTa shows the best performance. After 30 epochs of training, the training loss stands at 13.61, with the following metrics on the test set: accuracy at 0.92, precision at 0.68, recall at 0.72, and an F1-score of 0.70.

The performance across different entity classes highlights distinct patterns depending on the entity type. Traditional named entities such as *person names*, *official titles, times, places, object names,* and *materials*, show higher accuracy. Conversely, entity classes like *action causes* that involve a higher degree of interpretation or are abstract tend to show lower performance. Additionally, classes such as *deities* and *labors*, affected by data sparsity, also demonstrate weaker performance. Entities with moderate performance include action, quantitative entities (such as *cost*, *quantity*, and *measurement*), *object parts*, and *social categories*.

Future work will include continuous explorations of tagging schemas, optimization of dataset split, preprocessing methodologies including sentence segmentation and feature selection, and fine-tuning of hyperparameters. Following the training and evaluation of the SER task,

[1] CKIP BERT Base Chinese: https://huggingface.co/ckiplab/bert-base-chinese
[2] GuwenBERT: https://github.com/ethan-yt/guwenbert; a derived RoBERTa model: https://huggingface.co/KoichiYasuoka/roberta-classical-chinese-base-char
[3] SikuBERT: https://huggingface.co/SIKU-BERT/sikubert; SikuRoBERTa: https://huggingface.co/SIKU-BERT/sikuroberta

we will pursue two objectives: first, incorporating the model into our annotation platform, MARKUS; second, conducting experiments to integrate it with relation extraction and knowledge graph tasks. Moreover, we aim to extend the applicability and interoperability of our research beyond the confines of our current projects, contributing to the wider field of digital humanities, history, and infrastructure studies.