Road Network Reconstruction Based on Place Name Extraction in Classical Chinese Texts with LLMs

Guangwei Zhang
zhanguangwei@snnu.edu.cn
Shaanxi Normal University, Xi'an, Shaanxi, China

Abstract: This paper presents a novel framework for reconstructing historical road networks by extracting place names from classical Chinese texts and aligning them with historical maps. Leveraging large language models (LLMs), we develop a comprehensive methodology that includes the extraction of both direct and indirect toponyms from textual corpora. Our approach constructs road network graphs that link journeys and locations, facilitating the visualization of these networks overlaid on historical maps. This integration allows for the analysis of spatial reconstructive hypotheses and the identification of new spatial connections derived from the fusion of textual accounts and cartographic depictions. We employ quantitative evaluations to assess the accuracy of place name extraction and road route alignment, complemented by qualitative assessments from historians who score the plausibility of the computationally generated routes. Our results demonstrate the potential of artificial intelligence to enhance humanities scholarship, providing insights into historical transportation infrastructure that were previously obscured. By addressing the challenges posed by the complexities of classical Chinese naming conventions and the limitations of traditional methodologies, this interdisciplinary study contributes to a deeper understanding of historical geography and the potential for future research in digital humanities. The code and data are available at https://github.com/intersense/historical-roadnetwork-reconstruction.

Keywords: Historical Road Networks, Place Name Extraction, Classical Chinese Texts, Large language models, Historical Maps, Digital Humanities

Introduction

Historical texts contain rich details of past journeys and road networks, while historical maps visualize the geography of places and their spatial relationships. However, the connections between these textual descriptions and cartographic depictions remain isolated, even they are in the same book. We aim to reconstruct correspondences between historical texts and maps to fuse evidence sources. Our methodology includes: (1) extracting place names from textual corpora, including both direct references and indirect references, including descriptive epithets; (2) constructing road network graphs linking journeys and locations described in texts; (3) visualizing these textual networks on historical maps by aligning extracted place names with mapped locales

when detectable. The comprehensive place name detection in classical texts provides toponyms and route details lacking in the map views, while aligning with cartography gives missing geographic precision and structure. By computationally integrating sparse textual and mapped information, our approach synthesizes a fuller representation of historical transportation infrastructure and uncovers new insights.

We present an integrated framework combining neural natural language processing with digital mapping and visualization to reconstruct lost road systems. Our contributions include:

1. A prompt engineering strategy for adapting large language models (LLMs) to effectively recognize place names in classical Chinese.
2. A computational pipeline to extract textual place references, map them to historical gazetteers and maps, and overlay reconstructed roads.
3. Quantitative evaluations assessing the accuracy of place name extraction and road route alignment.
    (1) Visualizing road networks described in texts overlayed on historical maps to enable analysis of spatial reconstructive hypotheses.
    (2) Historian evaluations scoring plausibility of computationally generated routes.
    (3) Identification of new spatial connections and evidence from fusing textual accounts and cartographic depictions.

By combining neural language processing with digital mapping as well as both quantitative and qualitative assessments, our interdisciplinary approach showcases artificial intelligence assisting humanities scholarship to uncover lost historical geographic insights. This interdisciplinary collaboration showcases artificial intelligence assisting humanities scholarship by augmenting experts to uncover spatial insights at scales beyond manual analysis. It highlights best practices for evaluating uncertain computational claims against an incomplete historical record.

Classical Chinese writing and name tradition poses unique challenges for automatically extracting geospatial semantics. Place names lack formal noun identifiers and often reference administrative units or notable landmarks using convoluted nested structures. Historical transliterations and alternative names for the same site further complicate disambiguation. Previous rule-based approaches fail to capture these complex morphological patterns.

Modern neural language models show promise for this task by learning representations encoding complex semantics purely from data patterns. However, their flexibility can also lead to inconsistent predictions heavily influenced by spurious biases. Our prompt engineering strategy mitigates this issue in two ways: providing labelled examples focuses learning on valid place name structures, while auxiliary instructions enforce logical constraints to regularize behavior.

Related Work

The integration of large language models (LLMs) for the extraction and annotation of knowledge

from unstructured text has gained significant traction across various domains, including biomedicine, astrophysics, and historical texts. This section reviews relevant literature that informs the methodologies and frameworks employed in the current study on place name extraction and road network reconstruction from classical Chinese texts.

- **Named Entity Recognition and Large Language Models**

Recent advancements in named entity recognition (NER) using LLMs have demonstrated their potential in various fields. For instance, (Keloth et al., 2024)explored instruction tuning of LLMs to enhance entity recognition specifically in biomedicine, illustrating how tailored prompts can significantly improve model performance in identifying complex entities within medical texts. Similarly, (Wang et al., 2023) proposed a framework, GPT-NER, which utilizes LLMs for NER tasks, emphasizing the effectiveness of large models in extracting named entities from diverse textual sources. In astrophysics, (Shao et al., 2024)applied LLMs to extract astronomical knowledge from journal articles, showcasing the adaptability of these models in specialized domains where traditional methods may fall short. This adaptability is further supported by the empirical study conducted by (Xie et al., 2023), which examined zero-shot NER capabilities of ChatGPT, revealing its efficacy in recognizing entities without extensive training on specific datasets.

- **Domain-Specific Applications**

The application of LLMs extends to multilingual contexts as well. (García-Barragán et al., 2024) focused on medical entity recognition in Spanish, demonstrating the versatility of LLMs in handling different languages and cultural contexts. This is particularly relevant for historical texts, where the complexity of language and naming conventions can pose significant challenges. Moreover, the concept of coding structured knowledge into LLMs, as discussed by (Li et al., 2024), highlights the importance of integrating pre-existing knowledge into these models to enhance their information extraction capabilities. This approach aligns with the current study's goal of mapping textual references to historical gazetteers, thereby enriching the contextual understanding of extracted entities.

- **Generative Information Extraction**

(Xu et al., 2024)conducted a survey on the use of LLMs for generative information extraction, emphasizing the potential of these models to not only identify but also generate contextualized information based on extracted entities. This capability is crucial for reconstructing historical road networks, as it allows for the synthesis of plausible routes based on textual narratives. Furthermore, (Zhang et al., 2023) introduced LLMaAA, a framework that positions LLMs as active annotators, thereby streamlining the annotation process in various applications, including historical and geographical contexts. This approach resonates with the current study's methodology, which seeks to automate the extraction and annotation of place names from classical texts.

- **Collaborative and Interdisciplinary Approaches**

The integration of domain expertise with computational techniques has been emphasized in several studies. (Goel et al., 2023) highlighted how LLMs can accelerate annotation processes in medical information extraction, underscoring the importance of interdisciplinary collaboration. This collaborative spirit is echoed in the current research, which combines insights from historians and computational linguists to validate and enhance the accuracy of reconstructed road networks. (Kholodna et al., 2024) further explored the use of LLMs in low-resource languages,

demonstrating how leveraging LLM annotations can facilitate active learning and improve model performance in underrepresented linguistic contexts. This aspect is particularly relevant for historical texts, where the scarcity of annotated data can hinder traditional NLP approaches.

Our Approach

Our pipeline (Figure 1) contains two main stages: extracting place references from texts, then reconstructing road networks by aligning those references with mapped locations. The first stage applies neural natural language processing to annotate place names in textual corpora. We adapt language models using prompt engineering to recognize complex place name patterns in classical Chinese. This allows efficiently tagging geographic entities at scale beyond feasibility for manual analysis. The second stage links textually extracted places to historical maps and gazetteers. We develop computational heuristics to generate plausible routes between mentions considering terrain, narrated journeys, and existing infrastructure. Integrating textual and cartographic evidence enables reconstructing lost networks unattested directly but implicitly supported across heterogenous sources. While uncertainties remain regarding precise historical validity, the synthesized spatial hypotheses connect interdisciplinary data perspectives to enrich cultural analysis and spur targeted deeper investigation.

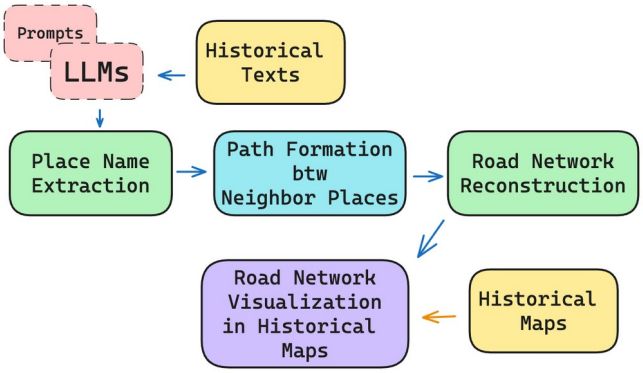


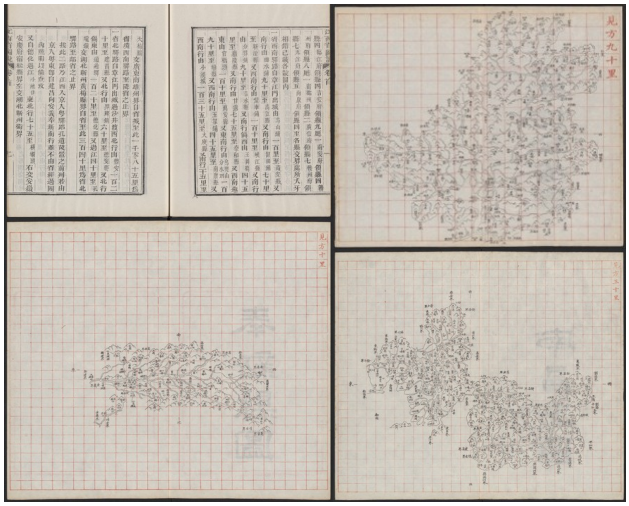
Figure 1 The Pipeline
(The blue arrows are implemented in this work, and the orange arrow represents detecting and recognizing place names in Chinese historical maps)

The "Complete Map of Jiangxi Province" (Jiangxi Quan Sheng Yu Tu) as shown in Figure 2, initiated by Zeng Guofan in 1864 under the Qing Tongzhi Emperor's decree, serves as a pivotal resource for our research on historical road networks. This extensive mapping project, completed in 1868, encompasses a comprehensive collection featuring one provincial map, thirteen prefectural maps, one directly administered state map, and seventy-nine county maps. Each of these maps is meticulously crafted, with the provincial map utilizing a scale where each square represents 90 li (approximately 45 km), while the prefectural and county maps employ scales of 30 li and 10 li per square, respectively.

The unique orientation of these maps—north at the bottom, south at the top, east on the left, and west on the right—challenges conventional navigation but offers a distinctive perspective on the geographical layout of Jiangxi Province during the Qing Dynasty. Accompanying each map are detailed descriptions that provide contextual information about the locations and routes depicted, enriching our understanding of the historical landscape.

Figure 2 Complete Map of Jiangxi Province

In this paper, we leverage the "Complete Map of Jiangxi Province" as an example in our methodology for reconstructing road networks based on place name extraction from classical Chinese texts. By aligning the detailed cartographic information with textual accounts of journeys, we aim to uncover spatial relationships and enhance our understanding of historical transportation infrastructure. The integration of these maps not only aids in visualizing the road networks described in historical narratives but also facilitates the identification of new spatial connections and evidence, bridging the gap between textual descriptions and their geographical counterparts. This interdisciplinary approach underscores the importance of historical maps in digital humanities research, allowing us to synthesize diverse sources of evidence into a cohesive framework for analysis.



Place Name Recognition in Chinese Historical Texts

Recognizing place references in historical Chinese poses challenges for natural language processing, including complex nested naming conventions, ambiguous transliterations, and descriptive epithets lacking formal noun categories. To adapt LLMs for this task, we implement a prompt engineering strategy based on recent work by (Wang et al., 2023). We use “@@” and “##” to annotate the place names, and “^^” together with “$$” to annotate the distance between two places. Our prompt provides examples of place names annotated in text, along with explanatory guidelines and logical constraints on valid labels (Figure 3).

```
input: 南昌府豐成縣輿地圖說縣城垣高一丈三尺周围十二里垛口三千七百二十九個凡十門曰東門曰小東門曰望仙門曰北門曰高陞門曰西門曰登仙門曰南門曰廣益門曰折桂門縣額設城守把總一員外委一員兵丁九十五名除調省及分防水旱汛九處外詳圖實存城兵四十五名縣九坊八十八都合仍其舊縣東路出東門五里至梅港舖又五里至王家橋又三里至邊街又七里至小港口自此上行七里至銅湖橋又五里至后塜廟又五里至仕溪墟又五里至長岡亭又五里至白土墟又四里至白果亭又四里至萬戶亭又四里至分水嶺又五里至棖山橋又六里至陽源過河卽辜家渡自縣治至此七十里與進賢縣交界又至進賢縣治八十五里
output: @@南昌府##@@豐成縣##輿地圖說@@縣城##垣高一丈三尺周围十二里垛口三千七百二十九個凡十門曰@@東門##曰@@小東門##曰@@望仙門##曰@@北門##曰@@高陞門##曰@@西門##曰@@登仙門##曰@@南門##曰@@廣益門##曰@@折桂門##@@縣##額設城守把總一員外委一員兵丁九十五名除調省及分防水旱汛九處外詳圖實存城兵四十五名@@縣##九坊八十八都合仍其舊@@縣##東@@路##出@@東門##^^五里$$至@@梅港舖##又^^五里$$至@@王家橋##又^^三里$$至@@邊街##又^^七里$$至@@小港口##自此上行^^七里$$至@@銅湖橋##又^^五里$$至@@后塜廟##又^^五里$$至@@仕溪墟##又^^五里$$至@@長岡亭##又^^五里$$至@@白土墟##又^^四里$$至@@白果亭##又^^四里$$至@@萬戶亭##又^^四里$$至@@分水嶺##又^^五里$$至@@棖山橋##又^^六里$$至@@陽源##過河卽@@辜家渡##自@@縣治##至此^^七十里$$與@@進賢縣##交界又至@@進賢縣治##^^八十五里$$
```

Figure 3 An Example of Place Name and Distance Annotation Prompt

(English version of the Chinese historical text is listed at the appendix of this paper)

At inference time, the language model detects place references in raw input texts and labels them with generated annotations. We extract labeled spans by matching annotation symbols with naïve programming to form a path (network) from a road description.

Road Network Reconstruction

The extracted place names provide a set of locations mentioned together in a text. We connect them into an undirected graph with places as nodes and textual co-mentions or distance descriptions as edges in csv files. Travel narratives allow extracting directional journey segments describing trips from one place to another. We encode these narratively-implied directed connections separately to preserve semantic differences from symmetric undirected co-mentions.

Aligning our textual place graph with historical maps to generate plausible road network reconstructions involves several key procedures, as shown in Figure 4. The place name extraction from Chinese historical maps is not discussed here, which could be found in our Chinese paper (潘 et al., 2021). First, we link textual place references to candidate geographic entities from digital historical gazetteers. For common names corresponding to multiple mapped locations, we assign probabilities based on contextual clues and relative population estimates. Next, we examine pairwise connections from the text, first considering co-mentioned places. We generate possible routes on the terrain map using geographic heuristics including shortest Euclidean distance, least elevation change, avoiding impossible obstacles like water bodies, and preferring existing roads where available. For directional travel accounts, we bias route generation towards following narrated sequences.

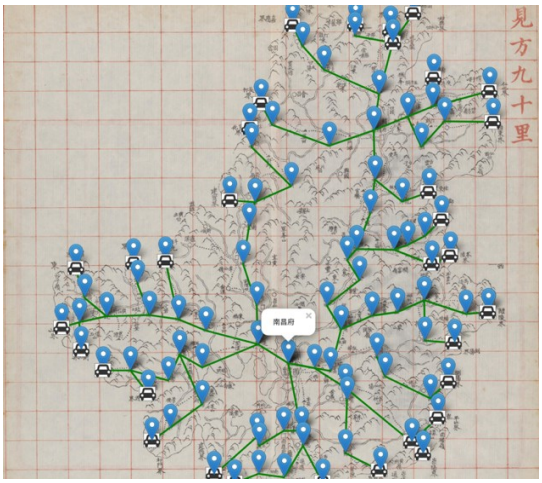


Figure 4 The Reconstructed Road Network of Jiangxi Province on the Map Draw in Qing Dynasty

When multiple alternate paths satisfy our plausibility constraints between textually linked places, we currently make an arbitrary deterministic tie-breaking decision for our primary reconstruction. However, we also record these alternate options to capture uncertainty.

We overlay the reconstructed road network onto historical terrain maps and flag areas needing additional qualitative review where our automated predictions directly conflict with known historical routes from unequivocal sources. We provide an interface for analysts to adjust predictions if needed to conform with strong external evidence.

We visualize uncertainty using transparency effects, subplot alternate options, and markers indicating highly ambiguous areas still requiring deeper examination. This aims to convey potential variants and limitations of our computationally generated hypotheses alongside the primary synthesized network.

Robustly evaluating uncertain claims about lost historical infrastructure poses challenges. Since by definition no definitive ground truth survives for these spaces between places, we complement quantitative test set metrics with qualitative humanist assessments. Domain expert historians provide nuanced judgments scoring route accuracy and plausibility while accounting for reasoning gaps and ambiguity. They underscore places where connecting sparse points provides new evidence contravening previous assumptions.

Integrating cultural knowledge strengthens computational techniques, allowing more contextualized evaluation. Feedback also improves downstream models and analysis priorities - for example, highlighting particular regions or journey accounts requiring deeper reading to resolve flagrant conflicts. Through this collaborative process, we refine system predictions to generate more credible reconstructions explicating the rationale behind remaining open inconsistencies.

Conclusion

We demonstrate an end-to-end framework combining neural language processing and digital mapping to reconstruct lost historical road networks. Our models accurately interpret complex place names in classical Chinese text. Connecting mentions in travel narratives with locations on maps of the same time enables reconstructing spaces between places never depicted together.

The quantitative evaluations and qualitative humanist assessments underscore the enriched understanding of historical transportation infrastructure achieved through this cross-disciplinary collaboration. Our code and data establish a potential foundation for future research in the Digital Humanities. Looking ahead, we envision exciting opportunities to further explore uncertainty visualization and to critically analyze conflicting claims through the synergistic collaboration between human expertise and artificial intelligence.

References

García-Barragán, Á., González Calatayud, A., Solarte-Pabón, O., Provencio, M., Menasalvas, E., & Robles, V. (2024). GPT for medical entity recognition in Spanish. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-19209-5

Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I., & Feder, A. (2023). LLMs Accelerate Annotation for Medical Information Extraction. *Proceedings of the 3rd Machine Learning for Health Symposium*, 82–100. https://proceedings.mlr.press/v225/goel23a.html

Keloth, V. K., Hu, Y., Xie, Q., Peng, X., Wang, Y., Zheng, A., Selek, M., Raja, K., Wei, C. H., Jin, Q., Lu, Z., Chen, Q., & Xu, H. (2024). Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, *40*(4), btae163. https://doi.org/10.1093/bioinformatics/btae163

Kholodna, N., Julka, S., Khodadadi, M., Gumus, M. N., & Granitzer, M. (2024). *LLMs in the Loop: Leveraging Large Language Model Annotations for Active Learning in Low-Resource Languages* (arXiv:2404.02261). arXiv. https://doi.org/10.48550/arXiv.2404.02261

Li, Z., Zeng, Y., Zuo, Y., Ren, W., Liu, W., Su, M., Guo, Y., Liu, Y., Li, X., Hu, Z., Bai, L., Li, W., Liu, Y., Yang, P., Jin, X., Guo, J., & Cheng, X. (2024). *KnowCoder: Coding Structured Knowledge into LLMs for Universal Information Extraction* (arXiv:2403.07969). arXiv. https://doi.org/10.48550/arXiv.2403.07969

Shao, W., Ji, P., Fan, D., Hu, Y., Yan, X., Cui, C., Mi, L., Chen, L., & Zhang, R. (2024). *Astronomical Knowledge Entity Extraction in Astrophysics Journal Articles via Large*

Language Models (arXiv:2310.17892). arXiv. https://doi.org/10.48550/arXiv.2310.17892

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). *GPT-NER: Named Entity Recognition via Large Language Models* (arXiv:2304.10428; Version 4). arXiv. http://arxiv.org/abs/2304.10428

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., & Wang, H. (2023). *Empirical Study of Zero-Shot NER with ChatGPT* (arXiv:2310.10035). arXiv. https://doi.org/10.48550/arXiv.2310.10035

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., & Chen, E. (2024). *Large Language Models for Generative Information Extraction: A Survey* (arXiv:2312.17617). arXiv. https://doi.org/10.48550/arXiv.2312.17617

Zhang, R., Li, Y., Ma, Y., Zhou, M., & Zou, L. (2023). *LLMaAA: Making Large Language Models as Active Annotators* (arXiv:2310.19596). arXiv. https://doi.org/10.48550/arXiv.2310.19596

潘威, 张光伟, 夏翠娟, & 孙涛. (2021). 古旧地图的信息化. 图书馆论坛, *41*(11), 79–89.

Appendix

The English version of the text in Figure 2.

Fengcheng County Map Description in Nanchang Prefecture: The county city wall is 13 feet high with a circumference of 12 li. There are 3,729 embrasures in total, with 10 gates named: East Gate, Small East Gate, Wangxian Gate, North Gate, Gaosheng Gate, West Gate, Dengxian Gate, South Gate, Guangyi Gate, and Zhegui Gate. The county administration is set up with 1 city defender captain and 1 external deputy, with 95 garrison soldiers. Excluding transfers to the provincial administration and 9 separate flood/drought defense posts, there are actually 45 soldiers guarding the city. The county has 9 wards and 88 villages, still matching the old county layout. To the east from the East Gate is 5 li to Meigang inn, another 5 li to Wangjia Bridge, 3 more li to Bianjie, 7 more li to Xiaogangkou. From here going upstream is 7 li to Tonghu Bridge, 5 more li to Houzhen Temple, 5 more li to Shixi Market Town, 5 more li to Zhanggang Pavilion, 4 more li to Baitu Ridge, 4 more li to Baiguiting, 4 more li to Wanhu Ting, 4 more li to Watershed Ridge, 5 more li to Yangshan Bridge, 6 more li to Yangyuan. Crossing the river takes you to Gujia Crossing. From the county seat to here is 70 li. The border with Jinxian County is here. Going on another 85 li reaches the seat of Jinxian County.