Detecting Entity Descriptions from Chinese Historical Texts

Ye Xia[1,†], Bin Wang[2,†], Linxuan Yu[3,†], Xiaoci Lin[1,†] and Hui Li[2,*,†]

[1]College of Artificial Intelligence, Nanjing Agricultural University, 210095, Nanjing ,China
[2]College of Humanities and Social Development, Nanjing Agricultural University, 210095, Nanjing ,China
[3]College of Sciences, Nanjing Agricultural University, 210095, Nanjing ,China

Abstract

Driven by an increasing number of digitized historical documents in machine-readable formats, researchers from various disciplines actively participate in the information extraction and exploration of historical documents, especially the recognition and classification of named entities in large-scale texts. Most existing studies focus on the identification of flat entities, however, the nested structures inside entities are often overlooked. In this paper, we focus on the extraction of nested entities in Chinese local gazetteers spanning over 8 centuries. We first propose an annotation guideline for two entity types and five entity categories in local gazetteers, which can be easily adapted to other domains. Then we utilize three popular span-based NER approaches in the context of Chinese historical texts, and analyze the corresponding results. Our preliminary study can enhance the existing geographical resources with entity information and be a reference for similar tasks within the field of digital humanities.

Keywords

Chinese historical texts, nested entity, span-based NER

1. Introduction

Named entity recognition (NER), which plays an important role in the area of natural language processing (NLP), identifies entities such as person, organization and location names from texts. Currently NER task has achieved a remarkable performance on texts written in modern languages. However, historical texts are still faced with multiple challenges, such as lack of resources, input noisiness, and domain heterogeneity [1]. Although transformer-based NER techniques have already been used on historical texts [2], fine-grained NER, e.g., nested entity recognition, has not been widely studied, especially for Chinese historical texts. The most widely used NER-flat approach on Chinese historical texts is the sequential labeling method "BERT-BiLSTM-CRF".

Chinese local gazetteers (also known as "difangzhi"), are historical records that contain comprehensive information about administrative units in China over time. In this study, we are particularly interested in extracting fine-grained entity information from large-scale Chinese historical texts. We make our efforts to extract the flat and nested entity mentions, e.g., local products, books and locations, from a sizable number of local gazetteers spanning over 8 centuries, using a computational approach.

2. Methodology

Within the scope of Chinese local gazetteers, our major focus is about two entity types, namely flat and nested entities. Five categories within these two entity types are defined and labeled with different

GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland
*Corresponding author.
†These authors contributed equally.
19220124@stu.njau.edu.cn (Y. Xia); 2022110023@stu.njau.edu.cn (B. Wang); 23121215@stu.njau.edu.cn (L. Yu); 19222119@stu.njau.edu.cn (X. Lin); 2021005@njau.edu.cn (H. Li)
0009-0000-3244-4069 (Y. Xia); 0009-0000-7685-3251 (B. Wang); 0009-0005-0961-5452 (L. Yu); 0009-0004-2972-2071 (X. Lin); 0000-0001-7050-1845 (H. Li)

tags: PRO for local product names, PER for person names, LOC for location names, BOK for book names, and TIM for temporal expressions.

Let a sentence $s \in S$ be denoted as a sequence of tokens $s = \{w_1, w_2, ..., w_n\}$, where w_i denotes the i^{th} token in the sequence. Let C denote a set of pre-defined categories $C = \{c_1, c_2, ..., c_n\}$. The goal of our task is to predict a list of tuples $T = \{< I_1^{head}, I_1^{tail}, c_1, d_1 >, < I_2^{head}, I_2^{tail}, c_2, d_2 >, ..., < I_m^{head}, I_m^{tail}, c_m, d_m >\}$, each of which refers to an entity mentioned in the sentence. I_i^{head} and I_i^{tail} represent the head index and the tail index of the i^{th} entity mention. c_i is the predicted entity category and d_i corresponds to the depth of this mention. m is the number of entity mentions detected within the given sentence. In this study, we consider the flat entity as a special case of the nested entity with the entity depth of zero (see Figure 1).

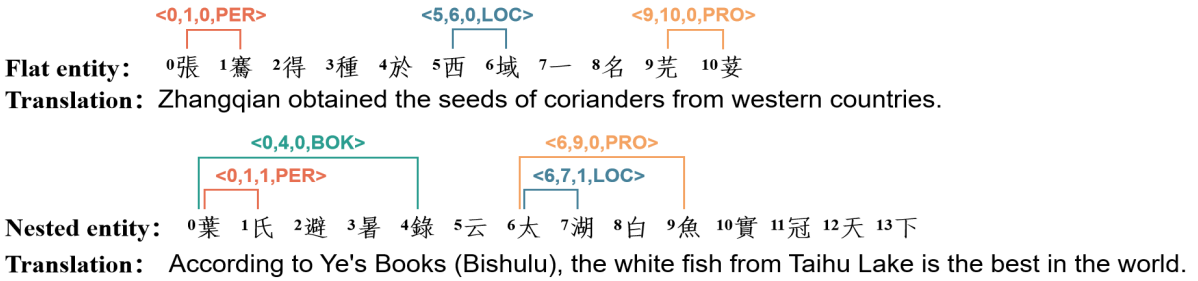

Figure 1: An example of the flat and nested entities annotated in Chinese local gazetteers.

Flat entities are annotated for all entity categories and nested entities are annotated within the categories of persons, locations and books in order to achieve a fine-grained entity detection. For instance, the category LOC covers the geographical names of a certain place, such as village, town or city, "高郵狀元墩" (Gaoyou Zhuangyuandun), the entity mention contains an embedded county name "高郵" (Gaoyou), so we consider this location name as a **nested** entity mention.

Our task here is to find all occurrences of the entities that belong to the categories indicated above, and use pre-defined tags to mark the beginning, the end, and the nested structure of each mention span. It should be noted that the tag sets not only refer to the full name of an entity, but also to the specific features embedded in a given entity. For example, the nested entity "波斯橙" (Bosi Cheng, Persian Orange) , the tag LOC is used inside the mention to capture the latent geographic feature (波斯, Persian) of this local product. We ask domain experts to manually assigned tags to each entity mention in texts and consult with each other or refer to external resources in case of entity ambiguities or uncertainties.

Considering the limited data scale, we prefer to fine-tune the pre-trained BERT-based model for NER task on classical Chinese texts instead of training our own from scratch. The span-based methods have advantages in easily identifying nested entities in different sub-sequences, therefore we intend to tackle the NER-nested task with three popular span-based approaches, namely MRC, Global Pointer and BERT-span, respectively.

- MRC. Li et al.[3] formulated the NER task as a Machine Reading Comprehension (MRC) task and transformed the tagging-style annotated dataset to a set of tuples {question, answer, context} to tackle nested entity problem.
- Global Pointer. Su et al.[4] leverages the relative positions through a multiplicative attention mechanism to identify the nested entities.
- BERT-Span. This approach leverages the strengths of BERT and span-based strategy to tackle the complexity of NER-nested task. This approach can effectively identify and extract the hierarchical relationships between nested entities.

3. Experiment and Evaluation

In this study, we use a digital collection of Chinese local gazetteers from the 12th to 20th century [5]. After text correction of misspellings, integration of metadata and manually annotation of entities, our

dataset contains 25,353 items of product descriptions and 940,189 entity labels. Table 1 shows the entity distribution of our dataset and we divide the dataset into three subsets, i.e., training, development and test set, with a ratio of 7:2:1. Apart from generic location names, we notice that there is a large proportion of nested entities containing ″LOC″ labels inside, which correspond to the ″Geo-related″ in Table 1.

Table 1
Entity Statistics for Our Dataset

Type	Category	Vocabulary Size	Geo-related
flat	PRO	10,423	✓
	LOC	6,445	✓
	BOK	102	✓
	PER	1,083	–
	TIM	540	–
nested	PRO	4,783	68.59%
	LOC	2,349	✓
	BOK	97	82.35%

We investigate the state-of-the-art (SOTA) pre-training language models for classical Chinese, and we find that BERT-ancient-Chinese is the best fit for our task since it outperforms others [6]. We fine-tune it on our annotated gazetteer dataset for the NER task with three span-based approaches, respectively. Table 2 illustrates the precision, recall and F1 values of entities in five categories using different span-based methods, and macro-average is calculated over all categories. We use bold to mark the highest value of each category. According to Table 2, it seems that Global Pointer outperforms others on macro-average scores and MRC outperforms others on identification of entities from PRO, LOC and PER categories.

Table 2
P, R, F1-Score, and Macro-average Results of Each Entity Category using Different Span-based Approaches

Methods	Category	P (%)	R (%)	F1 (%)
BERT-Span	PRO	81.22	82.34	81.77
	LOC	86.93	82.89	84.86
	BOK	83.97	86.29	85.12
	PER	82.17	85.87	83.98
	TIM	85.16	80.5	**82.77**
	Overall	83.89	83.58	83.7
Global Pointer	PRO	83.16	84.1	83.63
	LOC	86.86	81.62	84.16
	BOK	89.74	84.06	**86.81**
	PER	84.14	81.23	82.66
	TIM	81.51	81.41	81.46
	Overall	85.08	82.48	83.74
MRC	PRO	83.26	86.42	**84.81**
	LOC	86.53	85.37	**85.95**
	BOK	83.42	83.22	83.32
	PER	89.25	88.76	**89.00**
	TIM	81.54	80.43	80.98
	Overall	82.26	79.88	81.03

4. Conclusion

In this study, we focus on the nested entity extraction from large-scale Chinese local gazetteers. We utilize three popular span-based approaches with fine-tuning BERT-ancient-Chinese on our domain-specific dataset and the corresponding experimental results show the effectiveness and feasibility of span-based NER on Chinese historical texts. The extracted entity mentions in this ongoing study can

enrich the existed geographical resources with historical location names and local products. Our further step will be the entity linking with external resources, which will facilitate domain experts in the interpretation and understanding of the fine-grained knowledge embedded in the historical texts.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Chinese Central Universities (SKQY2022003) and the National Key Research Project on Rare-Book Collections (22GJK004). The authors are especially thankful to Prof. Dr. Ping Bao for the insightful comments, administrative technical support, and materials used for experiments. We are also grateful to Mr. Shun Zou for individual effort and encouragement to the successful completion of this paper.

References

[1] Won, M., Murrieta-Flores, P. and Martins, B. (2018). Ensemble named entity recognition (NER): evaluating NER tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, 5, p.2. doi: 10.3389/fdigh.2018.00002.

[2] Abadie, N., Carlinet, E., Chazalon, J. and Duménieu, B. (2022). A benchmark of named entity recognition approaches in historical documents application to 19th century French directories. *International Workshop on Document Analysis Systems*, pp. 445-460. Cham: Springer International Publishing. doi: 10.1007/978-3-031-06555-2_30.

[3] Li, X., Feng, J., Meng, Y., Han, Q., Wu, F. and Li, J. (2020). A unified MRC framework for named entity recognition. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5849–5859. 10.18653/v1/2020.acl-main.519.

[4] Su, J., Murtadha, A., Pan, S., Hou, J., Sun, J., Huang, W., Wen, B. and Liu, Y., 2022. Global pointer: Novel efficient span-based approach for named entity recognition. arXiv preprint. https://doi.org/10.48550/arXiv.2208.03054.

[5] Li, Y. and Li, H. (2022). Exploring the Rice Cultivars in Large-Scale Chinese Local Gazetteers: A Computational Approach. Plants, 11(23), p.3403. https://doi.org/10.3390/plants112334

[6] X. Hu, H. Zhang and Y. Sun. Chinese medical short text matching model based on fine-tuning BERT-Attention-BiLSTM. (2023). *23rd International Conference on Computer and Information Science*, pp. 91-96. doi: 10.1109/ICIS57766.2023.10210224.