A Geo-Tagged Classical Chinese Poetry Corpus

Weili Zhang[1], Xianpei Han[1], Le Sun[1], Ben He[2]
[1]Institute of Software, Chinese Academy of Sciences, Beijing China
[2]University of Chinese Academy of Sciences, Beijing, China
[1]{weili, xianpei, sunle}@iscas.ac.cn
[2]benhe@ucas.ac.cn

Abstract

The Chinese poetic tradition is the largest and longest continuous tradition in world literature. Classical Chinese poetry can be divided into certain standard periods or eras, in terms both of specific poems as well as characteristic styles. The knowledge about ancient civilizations can be learned from literature study on these poetry texts. However, there is little research focusing on building classical Chinese poetry resources for automatic natural language processing. In this paper, we take a preliminary step towards the above target and construct a geo-tagged Chinese poetry corpus. Specifically, an annotation criterion is first given to guide the tagging process for consistent annotation. Then we present details about the collecting, annotating and statistics about the data, from which a geo-tagged corpus of 5000 Chinese poems is built. Finally, the corpus is utilized to generate a geographic visualization, verifying its effectiveness on ancient civilization knowledge mining. Our corpus also provides a valuable resource for literature study, intelligent education, spatial data analysis, etc.

Keywords: Chinese poetry, geo-tagged poetry corpus, annotation criterion

1. Introduction

The Chinese poetic tradition is the largest and longest continuous tradition in world literature, which has a long history of more than 2,000 years. Classical Chinese poetry can be divided into certain standard periods or eras, in terms both of specific poems as well as characteristic styles. In ancient China, poetry is one of the most well-known and popular forms of literature, and nearly all scholars aspired to master poem composition. These classical poems help people to express their personal emotion, ambitions and thoughts. It also provides valuable literary texts for knowledge mining of ancient civilizations.

However, there was little research that focuses on building classical Chinese poetry resources for automatic natural language processing. Currently, most existing data sources about classical Chinese poetry are just raw texts. Due to the large size of classical Chinese poetry, and the genre diversity between ancient and modern Chinese language, it is difficult to analyze poetry texts using current natural language processing tools. To automatically understand a poem, a computing system must be able to extract different information about it, such as geographical locations, temporal information, related people and imagery in the poetry. Furthermore, all poems are correlated with each other based on different attributes, such as locations, poets, imageries, etc.

One of the most important knowledge for poem understanding is its geographic location. The geographic information provides background information where a poem was written, and the location itself provides comprehensive background about a poem. Furthermore, geographic information makes it easier to mine knowledge of authors, dynasties, and civilizations, by providing an important dimension for information integration, fusion and visualization. A simple example about the geo-labels of poetry is given in Figure 1.

In this paper, we construct a geo-tagged Chinese poetry corpus for automatically knowledge mining. Specifically, an annotation criterion is first given to guide the tagging process for consistent annotation. Then we present details about the collecting, annotating and statistics about the data, from which a geo-tagged corpus of 5000 Chinese poems is built. Finally, the corpus is utilized to generate a geographic visualization, verifying its effectiveness on ancient civilization knowledge mining. Our corpus also provides a valuable resource for literature study, intelligent education, spatial data analysis, etc.

2. Annotation Criterion

Classical Chinese poetry is written in ancient Chinese language, which is quite different from modern Chinese. Furthermore, China has a long history and most classical Chinese poems are written in ancient time, such as Qin dynasty (221–207 BC), Tang dynasty (AD 618–907) and Song dynasty (AD 960-1279). The above characteristics raise a number of unique challenges for geo tagging of classical Chinese poems. The main challenges are summarized as follows.

1) *Temporal variety of location names*. During the long history, many location names are changed. Therefore, it is quite common that the same location has different names in different times. For example, Soochow(苏州) has ancient names such as Gusu(姑苏), Wujun(吴郡) and WuXian(吴县);
2) *Location Name ambiguity*. Many location names are ambiguous. That is, the same location name may refer to different locations. For example, Han(韩) may refer to Han (state) or Han (Western Zhou state) in different dynasties. Therefore, the geo-tagging must distinguish ambiguous location names for down-stream applications.
3) *Location Role*. In classical Chinese poetry, some locations indicate where the poetry was written, while others are just mentioned in text, like the Hanshan Temple in Figure 1.

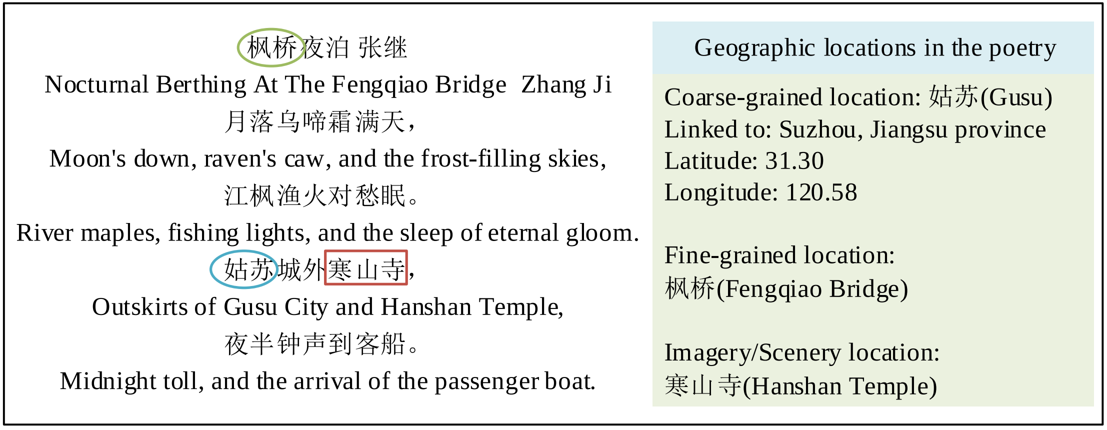
Based on the above observation, we geo-tag classical Chinese poems as follows:

1) Locations in poems are divided into two categories, one indicates where the poetry was written (e.g. Fengqiao and Gusu in Figure 1) -- *writing location*, and the other indicates locations mentioned in poetry text (e.g. the Hanshan Temple in Figure 1) -- *mention location*. Given a poem, annotators should label both categories of locations.
2) If more than one writing locations are annotated in a poem, *Part-Of* relations between them must be annotated. For example, in Figure 1, Fengqiao and

Gusu are both writing locations, then annotators must annotate that Fengqiao is *Part-Of* Gusu.

3) If a poem only contains one specific place, annotators should also give the city and the province this place belonging to by exploring the poem's context or background introduction. For example, given the verse 牧童遥指杏花村(The shepherd boy points at Xinghuacun), 杏花村 (Xinghuacun) must be annotated as a place, and the city it belongs to is Chizhou, and the province it belongs to is Anhui province.

4) All aliases, allusions of locations and place names should be labeled.



枫桥夜泊 张继
Nocturnal Berthing At The Fengqiao Bridge Zhang Ji
月落乌啼霜满天，
Moon's down, raven's caw, and the frost-filling skies,
江枫渔火对愁眠。
River maples, fishing lights, and the sleep of eternal gloom.
姑苏城外寒山寺，
Outskirts of Gusu City and Hanshan Temple,
夜半钟声到客船。
Midnight toll, and the arrival of the passenger boat.

Geographic locations in the poetry

Coarse-grained location: 姑苏(Gusu)
Linked to: Suzhou, Jiangsu province
Latitude: 31.30
Longitude: 120.58

Fine-grained location:
枫桥(Fengqiao Bridge)

Imagery/Scenery location:
寒山寺(Hanshan Temple)

Figure 1: A Chinese poem and the related geographic locations in it

5) To resolve the temporal variety problem of location names, all tagged ancient locations should be linked to its corresponding modern ones. This process is based on a mapping gazetteer between ancient and modern Chinese locations.

6) We use the BMEO annotation schema, where B, M, E, O correspondingly indicate begin, middle, end and out of a location name. For example, 姑/WB 苏/WM 城/WE 外/O 寒/IB 山/IM 寺/IE, where W and I represent the writing location and the mention location respectively. For each location mention, we also annotate its details, including its complete hierarchical district name, latitude and longitude.

3. Corpus

3.1 Data Collection

We prepare a comprehensive collection of poetry for geo-tagging. Specifically, we crawled data from two poetry websites -- Souyun[1] and Gushiwen[2]. After data integration and duplication clean, a total of 750 thousand poems are obtained, which stretch from Pre-Qin period to Qing dynasty, and poems of diverse types and genres are included. Table 1 shows statistics of this poetry collection.

3.2 Data Annotation

In order to construct a representative geo-tagged classical Chinese poetry corpus, we perform a poem selection step for each poetry genre. Specifically, all poems that have explanation notes are selected, because these notes provide background for geo-tagging. Because only a small part of poems have explanation notes, we also randomly sample poems without notes for a balanced distribution, and the sample size is the same as that of the noted ones. Finally, totally 14 thousand poems are selected, and they are reshuffled before provided to annotators.

For each poem, three annotators majoring in literature are invited for labeling according to the criterion in Section 2. Besides poems, annotators are also provided with online resource of poetry allusions[3] for reference, as well as a name mapping gazetteer[4] between ancient and modern locations for linking location names to its current locations. The final tagging results are determined by majority voting from three annotators' results.

3.3 Data Statistics

Finally, our geo-tagged corpus contains 5000 poems. Among them 2500 poems have geo-labels, and about 80.9% and 61.9% have writing locations and mention locations respectively, and 33.3% have both types of locations.

Besides, Figure 2 shows the top 5 ancient places possessing the most poetry, as well as the top 5 ancient places mentioned most frequently by poetry in the corpus. It's not surprising to see that some ancient capitals are so popular that poets love to writing for these places, even when they weren't staying there.

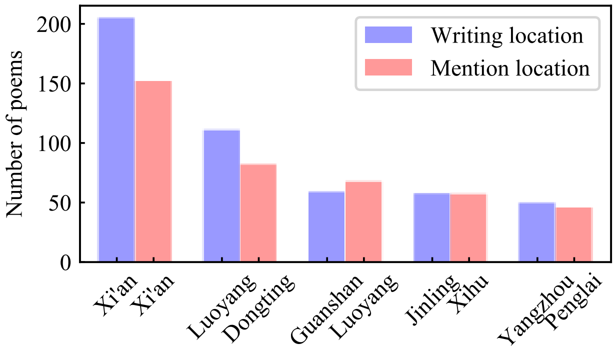


Figure 2: The top 5 writing locations and mention locations in poems of our corpus

[1] https://sou-yun.com/QueryPoem.aspx
[2] http://www.gushiwen.org/shiwen/
[3] http://cls.hs.yzu.edu.tw/orig/home.htm
[4] http://www.360doc.com/document/17/0902/01/11147672_684051406.shtml

4. Applications

The geo-tagged Chinese poetry corpus can be useful in many tasks. Basically, the corpus can provide training data for location extraction from ancient poetry text, which will further facilitate the automatically information extraction from these texts. Furthermore, our geo-tagging corpus provides valuable data resources for literature research, such as author profile, spatial text analysis, intelligent education, data visualization, etc.

We also demonstrate a simple application based on this corpus in Figure 3, which shows the top 11 places where authors write poems about Xi'an, the capital of Tang dynasty. We can see that, the capital of Tang dynasty has important influence on ancient poets: many poems are even from a foreign city -- Tokmak, as shown in Figure 3.

Types and genres	Shi Jing	Quatrain	Regulated verse	Iambic	Drama	Others
Total number	305	188,594	297,746	84,614	9,036	171,461
Number with notes	305	1,400	1,971	1,583	305	1,710

Table 1: Basic statistics of the raw poetry collection

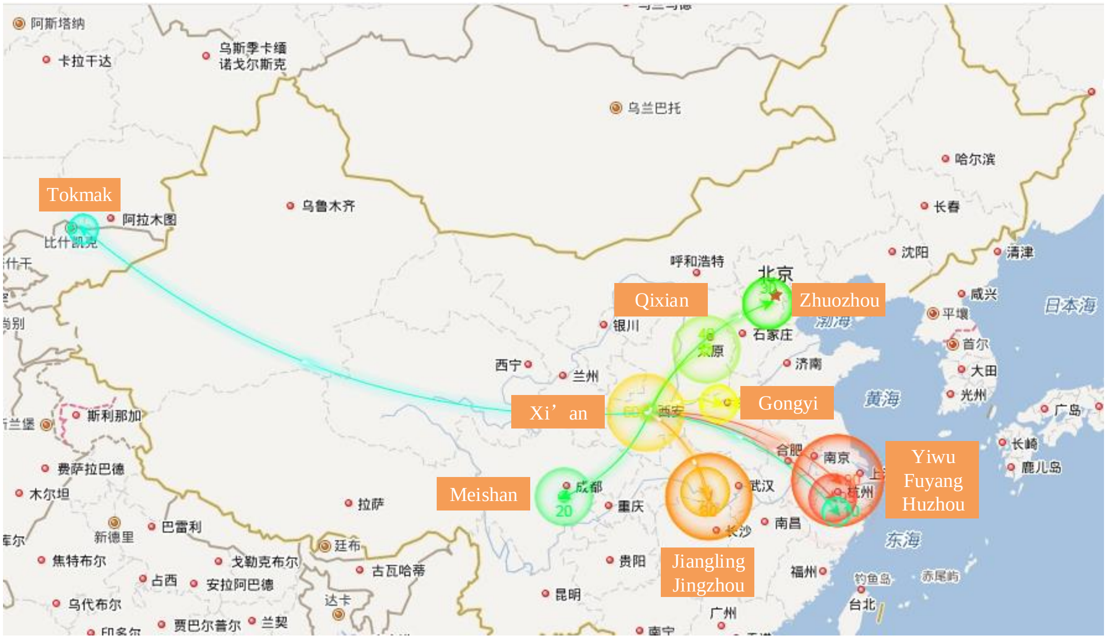


Figure 3: Top 11 places where authors wrote the most poetry using Xi’an as a mention location

5. Conclusions

The Chinese poetic tradition is the largest and longest continuous tradition in world literature. This paper builds a geo-tagged corpus, which provides a valuable resource for knowledge mining, literature research, spatial data analysis, and data visualization. Concretely, we design a basic annotation criterion, and 5000 classical Chinese poems are manually annotated.

6. Acknowledgments

This work is supported by Projects of The Chinese Language Committee under Grants no. WT135-24. Moreover, we sincerely thank the reviewers for their valuable comments.

7. Bibliographical References

Su, J., Zhou C., Li Y. (2007). The Establishment of the Annotated Corpus of Song Dynasty Poetry Based on the Statistical Word Extraction and Rules and Forms. *Journal of Chinese Information Processing*, 21(2):52–57.