A Study of Person Entity Extraction and Profiling from Classical Chinese Historiography

Yihong Ma[†,§], Qingkai Zeng[†], Tianwen Jiang[†], Liang Cai[‡], Meng Jiang[†]

[†]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
[‡]Department of History, University of Notre Dame, Notre Dame, IN 46556, USA
[§]School of Finance, Shanghai University of Finance and Economics, Shanghai, China
yihongma97@gmail.com,{qzeng,tjiang2,lcai,mjiang2}@nd.edu

ABSTRACT

When historians are interested in demographic and social network information of historical actors in the early Chinese empires (841 BC–1911 AD), very few studies have been done on entity retrieval from classical Chinese historiography. The key challenge lies in the low resource of the language: deep learning requires large amounts of annotated data and becomes impracticable when such data is not available. In this study, we employ domain experts (history professors) to curate a set of *person* entities and their profile attributes (e.g., *courtesy name*, *place of birth*, *title*) and relations (e.g., *father-son*, *master-disciple*) from two books, *Records of the Grand Historian* and *Book of Han*. We develop a pattern-based bootstrapping approach to extract the information with a very small number (i.e., 1 or 2) of seed patterns. Experimental results show the effectiveness as well as the limitations of the iterative method. We would appreciate research of digital humanities to address the challenges in entity retrieval from low-resource languages.

CCS CONCEPTS

• **Information systems** → **Data mining**; *Information extraction.*

KEYWORDS

Information extraction, Entity profiling, Classical Chinese, Textual pattern, Bootstrapping

1 INTRODUCTION

Historians are interested in the classical Chinese literature as it witnessed the rise and fall of the early empires and dynasties [2, 3, 18, 32]. Currently, they have to spend a large amount of time reading those books and digging out *who came from where*, *who did what*, *who studied from whom*, and so on, and before it, they had to spend even longer time to learn the classical Chinese language even some of them are (absolutely modern) Chinese [11, 22, 26]. Therefore, the idea of utilizing digital technologies to *extract person entities and their profiling attributes from classical Chinese text* becomes promising and exciting in the community, as natural language processing (NLP) and entity retrieval have been developing and accelerating at an unprecedented speed today.

However, it is a rather challenging task due to lack of annotated data. Historians write papers, publish books, but rarely build entity databases about ancient China. As we know, NLP is being revolutionized by deep learning with neural networks. However, deep learning requires large amounts of annotated data, and its advantage over traditional statistical methods typically diminishes when such data is not available. How to address the issue of *low resource* in the task of entity extraction and profiling from classical Chinese text is still an open problem.

Meng Xi 孟喜		
	Our approach	**Truth**
Courtesy name	长卿	长卿
Hometown	東海蘭陵	東海蘭陵
Title(s)	郎, 丞相掾, 名之	郎, 丞相掾, 曲臺署長
Father	孟卿	孟卿
Son	N/A	N/A
Master(s)	田王孫, 同郡碭田王孫	田王孫
Disciple(s)	沛翟牧子兄, 同郡白光少子, 疏廣, 后蒼	翟牧, 白光, 趙賓, 焦延壽

Table 1: The task is to extract person entities and their profiling attributes from classical Chinese text. Our approach can find most of the attribute values correctly (i.e., marked by underlines), compared with the ground truth annotated by history professors.

In this study, we collect a ground-truth dataset for evaluating on the task, propose a pattern-based information extraction approach that requires very limited prior knowledge of classical Chinese, and conduct experiments to show its effectiveness and limitations.

First, we recruit domain experts (history professors) to annotate person entities and their profiles from two classical Chinese books, *Records of the Grand Historian* (authored by Sima Qian, completed in c. 86 BC) and *Book of Han* (authored by Ban Gu, completed in 111 AD). Table 1 gives an example of the annotated profile of Meng Xi 孟喜 (~90 BC–~40 BC). We focus on three attributes (i.e., *courtesy name*, *place of birth*, *positions/titles*) and two relations (e.g., *father-son*, *master-disciple*), because (1) these are main contents in the work of Chinese historiography and (2) historians are very interested in how the government mechanisms were influenced by family and master-disciple relationships in the ancient time. The domain experts generated fifty handcrafted patterns to extract the above information. They validated the attribute values and assessed the reliability of the patterns (see Table 4). Moreover, they annotated 15 complete person profiles (with 158 attribute values) that they feel the most interested in. This dataset can serve as

	Meng Xi 孟喜	Meng Qing 孟卿	Yan An Le 顏安樂	Zhang Yu 張禹
Courtesy name	長卿	N/A	公孫	子文
Hometown	東海蘭陵	東海	魯國薛	河內軹
Title(s)	郎, 曲臺署長, 丞相掾	N/A	齊郡太守丞, 大司農	郡文學, 光祿大夫, 東平內史
Father	孟卿	N/A	眭孟姊	N/A
Son(s)	N/A	孟喜	N/A	N/A
Master	田王孫	蕭奮	眭孟	施讎
Disciple(s)	趙賓, 白光, 翟牧, 焦延壽	后倉, 閻丘卿, 疏廣	泠豐, 任公, 冥都, 筦路	彭宣, 戴崇

Table 2: Four examples of profiles of the historical actors in the *Han Dynasty*. We focus on the three attributes and two relations.

ground truth for evaluating such information extraction methods on the classical Chinese literature.

Second, we propose a bootstrapping approach to extract person entities and profiles requiring very little prior knowledge of the language. The algorithm starts from only one or two simple seed patterns, finds the attribute values, and then use them to discover more complicated patterns. It has an estimator to access the reliability of patterns during the iterative process. So, the new attribute values extracted from more reliable patterns are more likely to be trustworthy and can be used to infer patterns in next iterations.

Experiments show that textual patterns achieve an F1 score of 0.851 on 15 ground-truth person profiles. Table 1 shows a comparison between the generated profile (left) and the ground-truth profile (right) of Meng Xi 孟喜. On the other side, the bootstrapping method achieves the highest performance after 7 iterations to find a set of related patterns and extract *person-title* pairs, while meeting barriers to find more patterns for other attributes and relations.

We summarize our contributions in this study as follows:

- **New dataset:** We recruit history professors to curate a set of person profiles from classical Chinese literature.
- **New approach:** We develop a bootstrapping method based on textual patterns to extract the person entities and attributed information, requiring little prior knowledge.
- **Effectiveness:** Experiments show that textual patterns make an F1 score of 0.851 on 15 person profiles annotated by the domain experts. Limitations are discussed in Section 4.3.

2 ENTITY EXTRACTION AND PROFILING

2.1 Person Entity Extraction

It sounds like a subtask of the standard named entity recognition (NER) task – it narrows down from recognizing multiple types of entities (i.e., persons, locations, organizations) to only one type. However, it has to face a challenge when put into the classical Chinese text: in classical Chinese literature, there are many different ways of mentioning a specific person. A person has first name, last name (family name), and courtesy name; and he is also recognized by his hometown and title/position in the government. For the sake of readability, let us take the President of United States Donald J. Trump as an example. "Donald" is his first name and *suppose* "John" (J.) can be considered as his courtesy name. (Ancient Chinese people do not have middle name. They have courtesy name.) He was born in New York. So, all the following could be used in the classical Chinese literature to mention President Trump:

- donald,
- trumpdonald,
- presidentdonald,
- trumpdonaldjohn,
- newyorktrumpdonald,
- newyorktrumpdonaldjohn.

Note that there would be no white-space (nor upper-case) to split the words. Complicated patterns have to be designed or recognized in the extraction methods.

Historiography Book	# Sentences	# Words
Records of the Grand Historian	32,564	615,457
Book of Han	40,114	874,165

Table 3: Statistics of the dataset (two books).

2.2 Person Entity Profiling

Given the classical Chinese historiography, the task of person entity profiling aims to extract demographic attributes (e.g., *courtesy name*, *place of birth*, *title*) and social relations (e.g., *father-son*, *master-disciple*) and to generate a complete profile for the person entities extracted in Section 2.1.

Table 2 shows examples of the profiles of Confusians in the Han Dynasty such as Meng Xi 孟喜, Meng Qing 孟卿, Yan An Le 顏安樂, and Zhang Yu 張禹. Some of the attribute values are "N/A" if not available in the text corpora.

There are two typical challenges of this task. One is the variety of demographic attributes. Each type of attributes needs a set of specific, reliable extractors, which requires prior knowledge of the classical Chinese language. The other challenge is typically for the Chinese historiography: Zero Pronoun (ZP), which stands for pronouns that are omitted when they are pragmatically or grammatically inferable from the context. Here is an example taken from *Records of the Grand Historian*, where the ZPs (denoted as ϕ) all refer to "Mr. Chunshen" 春申君:

> [春申君] 者，ϕ 楚人也，ϕ 名歇，ϕ 姓黃氏。ϕ 游學博聞，ϕ 事楚頃襄王。
>
> (Translation: [Mr. Chunshen], ϕ was born in Chu, ϕ's first name is Xie, ϕ's family name is Huang. ϕ travelled over the country and enriched his knowledge, ϕ served King Qingxiang of Chu.)

This sentence indicates three attributes (i.e., *hometown*, *first name*, *last name*) and one relation (i.e., *master-disciple*) about Mr. Chunshen 春申君. However, ZP makes it challenging to link the attributes and relations in the context with the person entity. Moreover, we observe that ZPs occur not only in the same sentence with the mention of the person entity but also across several sentences in the same paragraph. In biographical historiography, each chapter discussed the life story of a certain person. So, we adopt the following assumption (learned from history professors) to resolve

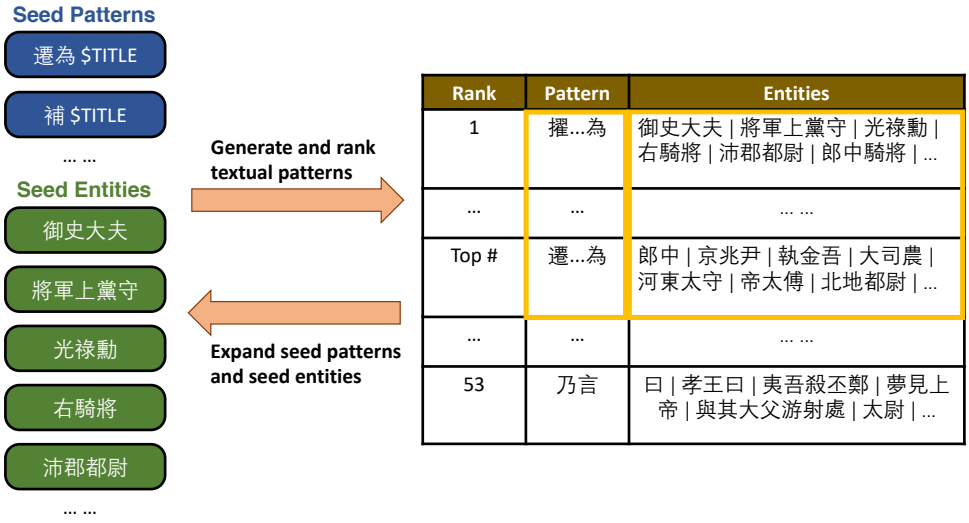

Figure 1: The diagram of the proposed pattern-based bootstrapping method.

the ZP issue: given a paragraph, as long as a person entity was extracted in the first clause, the ZPs in every clause of the paragraph refer to that person entity. This will help us propose an approach to extract person-attribute/relation pairs when the extractors were only able to find the attributes and relations in local contexts.

3 THE PROPOSED APPROACH

In this section, we first introduce how the dataset was curated with handcrafted patterns by domain experts. Next, we present a pattern-based bootstrapping method to find the entity information with a small number of seed patterns.

3.1 Data Curation with Handcrafted Patterns

We curate a dataset of two classical Chinese historiography books, *Records of the Grand Historian* (authored by Sima Qian, completed in c. 86 BC) and *Book of Han* (authored by Ban Gu, completed in 111 AD). Table 3 lists the statistics of the dataset.

3.1.1 Patterns for person entity extraction. The domain experts we recruited to annotate the data contribute the following patterns to recognize mentions of person entities:

- $FirstName,
- $LastName + $FirstName
- $Title + $FirstName,
- $LastName + $FirstName + $CourtesyName,
- $Hometown + $LastName + $FirstName,
- $Hometown + $LastName + $FirstName + $CourtesyName.

3.1.2 Patterns for entity profiling. Table 4 presents 50 textual patterns that were used to extract a set of candidates of person's attribute or relation values. Some attributes such as *hometown* and *father-son* have a small number of patterns. Some such as *title* and *master-disciple* have a large number of patterns. The domain experts also annotated whether the attribute values and relations are true or false. For each pattern, we give three numbers associated with its extractions:

- *#Values:* The number of (person entity, attribute or relation value)-pairs extracted by the pattern.
- *#True Values:* The number of true pairs annotated by the domain experts.
- *Reliability:* It describes whether a pattern is reliable for extracting true values. It is calculated as

$$Reliability = \frac{\#True\ Values}{\#Values}, \tag{1}$$

which gives a score between 0 and 1.

Specifically, for *person*, pattern [$Person 者] was the first pattern that the domain experts come up with. "者" is a typical symbol in classical Chinese that indicates the appearance of a person. The person entities extracted by pattern [$Person 者] may be in any of the 6 forms of person entity mentions in Section 3.1.1. Another frequent pattern is [$Person 字 $CourtesyName]. Unlike pattern [$Person 者], person entities extracted by pattern [$Person 字 $CourtesyName] strictly follow the form of *last name + first name*. It is a more reliable pattern. As the table shows, the reliability of pattern [$Person 字 $CourtesyName] is 1 and the reliability of pattern [$Person 者] is only 0.6087 though the number of extracted person-value pairs is smaller (205 vs. 299).

Most of the patterns for *hometown*, *father-son*, and *master-disciple* are highly reliable (higher-than-0.96 reliability), except patterns [, $Father 子] (ID 38) and [, 事 $Master] (ID 46) of reliability 0.8571 and 0.8356, respectively. Among the 28 patterns for attribute *title*, only 4 patterns have reliability of lower than 0.8 and only one has a reliability score of lower than 0.7, i.e., [至 $Title] (ID 36). Among all the 50 handcrafted patterns, 35 (70%) patterns have reliability score of 1; 5 (10%) patterns have reliability score of lower than 0.8.

3.2 Pattern-based Bootstrapping

We propose a new approach to extract person entities and profiles from classical Chinese historiography requiring very little prior knowledge of the language. Generally, it is an iterative method that uses textual patterns to extract attribute or relation values from text data. Figure 1 shows the diagram of one iteration in the

ID	Attribute	Pattern	Example	#Values	#True Values	Reliability
1	person	$Person 者	陳丞相平 者	299	182	0.609
2	person	$Person 字 $CourtesyName	王莽 字巨君	205	205	1.000
3	person	$Person，$Hometown 人也	申屠嘉，梁 人也	106	103	0.971
4	person	$Person，$Hometown 人	朝鮮王滿，燕 人	46	34	0.739
5	hometown	，$Hometown 人也	，陽城 人也	190	189	0.995
6	hometown	，$Hometown 人	，高陽 人	11	11	1.000
7	hometown	徙 $Hometown	自下邑徙平陵	16	16	1.000
8	courtesy name	，字 $CourtesyName	，字長卿	21	21	1.000
9	title	拜為 $Title	拜為上卿	22	22	1.000
10	title	拜 $Person 為 $Title	拜仁 為郎中令	8	8	1.000
11	title	遷 $Title	遷東平太傅	74	64	0.865
12	title	遷為 $Title	起遷為國尉	36	36	1.000
13	title	遷 $Person 為 $Title	遷廣明 為淮陽太守	1	1	1.000
14	title	遷至 $Title	稍遷至栘中廄監	19	19	1.000
15	title	封為 $Title	綰封為長安侯	18	18	1.000
16	title	封 $Person 為 $Title	孝景後三年封蚡 為武安侯	3	3	1.000
17	title	召為 $Title	復召為郎	2	2	1.000
18	title	召 $Person 為 $Title	於是上召寧成 為中尉	5	5	1.000
19	title	補 $Title	以選除補御史掾	41	40	0.976
20	title	察... 為 $Title	以郡吏察廉為樓煩長	8	8	1.000
21	title	舉為 $Title	後以御史舉為鄭令	8	8	1.000
22	title	舉... 為 $Title	復舉賢良為河南令	11	10	0.909
23	title	擢為 $Title	擢為光祿大夫	10	10	1.000
24	title	擢 $Person 為 $Title	因擢延壽 為諫大夫	3	3	1.000
25	title	徵為 $Title	徵為廄丞	14	11	0.786
26	title	徵 $Person 為 $Title	徵由 為大鴻臚	5	5	1.000
27	title	徙為 $Title	徙為頻陽令	11	11	1.000
28	title	徙 $Person 為 $Title	徙立 為太原太守	2	2	1.000
29	title	復為 $Title	後復為淮陽都尉	15	14	0.933
30	title	以 $Title 察	以郡吏 察廉為樓煩長	4	4	1.000
31	title	薦為 $Title	薦為議郎	4	4	1.000
32	title	薦 $Person 為 $Title	薦宣 為長安令	3	3	1.000
33	title	贖為 $Title	贖為庶人	8	8	1.000
34	title	立為 $Title	自立為代王	24	21	0.875
34	title	為 $Title	為駙馬都尉侍中	193	151	0.782
35	title	$Person 為 $Title	禹 為丞相史	45	32	0.711
36	title	至 $Title	至中大夫	115	37	0.322
37	father-son	，$Father 子也	，秦莊襄王 子也	25	24	0.960
38	father-son	，$Father 子	，文公 少子	14	12	0.857
39	father-son	，其父 $Father	，其父高祖中子	3	3	1.000
40	father-son	，父 $Father	，父號孟卿	6	6	1.000
41	father-son	$Son 父曰 $Father	悼侯 父曰隱太子友	18	18	1.000
42	master-disciple	從 $Master 受...	從太中大夫京房 受易	12	12	1.000
43	master-disciple	事 $Master 受...	又事前將軍蕭望之 受論語	2	2	1.000
44	master-disciple	$Master 授 $Disciple	常 授梁蕭秉君房	52	52	1.000
45	master-disciple	，授 $Disciple	，授翼奉、蕭望之、匡衡	25	25	1.000
46	master-disciple	，事 $Master	，事太傅夏侯勝	73	61	0.836
47	master-disciple	事 $Master 為 $Title	事梁孝王 為中大夫	3	3	1.000
48	master-disciple	弟子... 者，$Master	弟子遂之者，蘭陵褚大，東平嬴公	4	4	1.000
49	master-disciple	受... 於 $Master	嘗受韓子、雜家說於騶田生所	3	3	1.000
50	master-disciple	與 $Person 俱事 $Master	與顏安樂 俱事眭孟	6	6	1.000

Table 4: Patterns manually annotated by domain experts to find person entity profiles, where underlines mark the values extracted by the patterns.

pattern-based bootstrapping method. It starts with only one or two simple seed patterns for each attribute. Because the number of seed patterns is small, it would not take much effort to find one or two. For example, [遷為 $Title] (i.e., [relegated to $Title]) and [補 $Title] (i.e., [filled in the position of $Title]) were the two reliable seed patterns for the attribute *title*. The iterative method runs the following steps until convergence.

Step 1: Generating pattern candidates. Candidate patterns are generated using contextual features of the target value v_i in the clause. We find that target values are more likely to be at the end of the clause because of the linguistic structure. Therefore, the commonly used *skip-gram* contextual pattern "w_{-1}____w_1" [28] would not work for our task. Instead, we explore two different kinds of contextual features described as follows:

- **$Pattern $Value.** The textual pattern is a window of a certain size of Chinese characters before a target value. For example, if the target value is $Title, we can find the pattern candidate [遷為 $Title] (i.e., [relegate to $Title]), when the window size is 2.
- **$Pattern $Entity $Pattern $Value.** Both a window of one Chinese character before *$Entity* and all characters between *$Entity* and *$Value* are selected as the contextual feature. For example, if $Title is the target value and $Person is the entity that has already been extracted in Section 3.1.1, we can find a new pattern candidate [遷 $Person 為 $Title] (i.e., [relegate $Person to $Title]).

Step 2: Ranking pattern candidates. It is nontrivial to rank the quality of pattern candidates. It has two serious issues when considering all the unlabeled entities as false: (1) penalized reliable patterns that extracted true unlabeled values and (2) could not penalize unreliable patterns that extracted false unlabeled values. To address these issues, we define the estimation score of pattern reliability as follows:

$$r(p) = w_1 \cdot \frac{\sum_{v \in \mathcal{V}_p} \left(1 - \min_{v^+ \in \mathcal{V}^+} d(v, v^+)\right)}{\sum_{v \in \mathcal{V}_p} freq(v)} + w_2 \cdot \left(1 - \frac{\max_{v \in \mathcal{V}_p} freq(v)}{\sum_{v \in \mathcal{V}_p} freq(v)}\right) \in [0, 1],$$

where p is a textual pattern, v is a value string, v^+ is a true value string, $\mathcal{V}_p$ is the set of unique value strings extracted by pattern p, $\mathcal{V}^+$ is the set of unique true value strings; $d(v_1, v_2)$ is the normalized hamming distance between the two value strings, $freq(v)$ is the frequency of the value string v. w_1 and w_2 are weights: $w_1 + w_2 = 1$.

The estimator includes two kinds of features:

(1) *The textual similarity between the pattern's extracted values and true values:* If the value a pattern extracted is very similar with one true value, the value is likely to be true and the pattern is likely to be reliable. For example, suppose "Tai Shou 太守", the name of an official position, has been in the set of true values (as $Title). Then the value "Nan Yang Tai Shou 南陽太守" extracted by a pattern, which means the Tai Shou 太守 ruling a place called Nan Yang 南陽, is likely to be a good value (as $Title). We use *Hamming distance* as the metric to measure the similarity between two value strings. Hamming distance is defined as the minimum number of substitutions required to change one string into the other.

(2) *Variety of the pattern's extracted values:* A pattern would be more reliable if it extracted more true values. Besides the frequency, we try another measurement: we assume that if there was a value whose frequency dominates the set of values one pattern extracted, the pattern would be less reliable. So we use 1 minus the ratio of the count of the most frequent value over all the value counts.

Step 3: Selecting new patterns and extracting new values for the next iteration. For each pattern, we calculated the reliability score $r(p)$ and the frequency of values that it extracted. For the next iteration, we first filter out the patterns whose frequency is below a threshold and then select top patterns of the highest $r(p)$. After that, we expand the set of true values $\mathcal{V}^+$ by adding the values extracted by the new patterns.

4 EXPERIMENTS

In this section, we first evaluate the quality of handcrafted patterns given by domain experts. Then we evaluate the effectiveness of the bootstrapping method. Finally, we discuss the limitations.

4.1 Evaluating the Handcrafted Patterns

Here we conduct experiments to answer: do the handcrafted patterns extract correct person entities, attributes, and relations?

Evaluation methods. We use the 15 complete person profiles (with 158 values) as ground truth. We use standard Information Retrieval metrics: *Precision*, *Recall*, and *F1 score*. *Precision* is the fraction of true attribute or relation values (i.e., values that find a match in the corresponding attribute in ground-truth profiles) among all values extracted by handcrafted patterns. *Recall* is the fraction of true attribute or relation values among all ground-truth values. *F1 score* is the harmonic mean of *Precision* and *Recall*.

Evaluation results. Regarding the entity profiles, handcrafted textual patterns achieve a *Precision* of 0.901, a *Recall* of 0.803, and an *F1 score* of 0.851. Table 1 shows a comparison between the generated profile (left) and the ground-truth profile (right) of Meng Xi 孟喜, where most of the values extracted are correct. We also find the following limitations of the handcrafted patterns. First, the different forms of person entity mentions make entity linking (i.e., mention alignment) difficult. For example, in the *master-disciple* relation of Meng Xi 孟喜, "同郡白光少子", i.e., Bai Guang 白光 whose courtesy name is Shao Zi 少子 from the same ("同") hometown ("郡") as Meng Xi's, extracted by handcrafted patterns should refer to Bai Guang 白光 in the annotation. Second, ZP problem was resolved in most of the cases but may still assign attributes or relations to wrong persons. For example, in the *master-disciple* relation, Hou Cang 后蒼 and Shu Guang 疏廣 are indeed disciples of Meng Xi 孟喜's father Meng Qing 孟卿, but are mistakenly regarded as the disciples of Meng Xi 孟喜 due to the assumption.

4.2 Evaluating the Effectiveness of the Bootstrapping Method

We conduct experiments to see if the bootstrapping method can find the set of handcrafted patterns with only one or two seed patterns and see if the attribute values can be accurately extracted.

Parameter settings. We set the window size as 2. The frequency threshold of patterns is 10. The number of top patterns selected per iteration is 10. We run until convergence but just report the first 10 iterations for the sake of space. The weights of pattern reliability features are $w_1 = w_2 = 0.5$.

Evaluation methods. Here are the metrics for the two tasks.

Task of pattern extraction: We evaluate the performance on extracting patterns for the *title* attribute. We use the metric ***P**recision@K*, which is the fraction of top K scored generated patterns that are in the ground-truth pattern set. We also define a new metric ***C**overage@K* for the task, which is the fraction of top K scored ground-truth patterns that are extracted by the bootstrapping method. The generated patterns were scored by the reliability estimation in Step 2 in Section 3.2 and the ground-truth patterns were scored by the reliability in Table 4. Average precision (***AP***) computes the average precision value for coverage over 0 to 1.

Task of person-title pair extraction: We first assign a confidence score to each *person-title* pair by weighting the reliability score of the textual pattern that extracts *person* and *title* respectively. We evaluate the person-title pairs extracted by the bootstrapping method at different numbers of iterations with *Precision-Recall* curves. *Precision* is the fraction of true person-title pairs among all person-title pairs generated by handcrafted patterns. *Recall* is the fraction of true person-title pairs among all 516 ground-truth person-title pairs. *AUC* is the area under the curve.

Results on pattern extraction. The bootstrapping method had been improving the performance of pattern extraction since it started, while after certain iterations the performance turned to be worse. From Table 5, running the bootstrapping algorithm for 3 iterations can increase *AP* by 42.55%, compared to running only one 1 iteration. After around 5 iterations, *AP* displays a continuous trend of declining and iteration 10 gives the lowest *AP* of 0.131, which is a decrease of 44.26% from iteration 1. What's more, ***C**overage@K* no longer update after 7 iterations. It indicates that the bootstrapping may meet certain barriers in extracting more patterns.

After observing the result patterns, we can infer some limitations on pattern extraction of the pattern-based bootstrapping method:

First, there exist many patterns with either a relatively low frequency (i.e. less than 20) or lack of interpretability (i.e. patterns with scarcely any actual semantic meaning but somehow capable of extracting "good" entities, which are still considered "good" by our method) that tend not to be found by our domain experts, which we should be reasonably tolerant of.

Second, the pattern-based bootstrapping method is not good at abstracting the first type of contextual patterns mentioned in Step 1 in Section 3.2. Human experts can easily generalize patterns with a $v. + prep.$ structure that are composed of different verbs but the same pronoun into one super-category: *prep*. For example, it is reasonable for domain experts to find such common feature of patterns like [拜為 $Title], [擢為 $Title], [舉為 $Title] and etc., all of which mean [promote to $Title], and generalize them into pattern [為 $Title] (i.e., [to $Title]). However, the bootstrapping method tends not to capture such abstraction of patterns and therefore generates a subset of certain ground-truth patterns, which pulls down the evaluation metric.

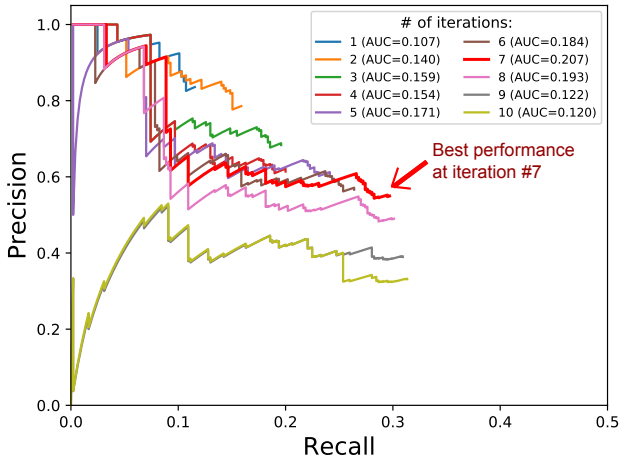

Figure 2: The performance of the bootstrapping method on *person-title* pair extraction gradually improved through iterations. *AUC* increased from 0.107 to 0.207 (in iteration 7) and decreased after the point.

Results on person-title pair extraction. Running the bootstrapping method for more iterations generally increases the performance of *person-title* pair extractions, while after certain iterations the performance starts to shrink. From Figure 2, *AUC* keeps increasing in the first 7 iterations, achieving a maximum of 0.207 in iteration 7, and then begins to decrease in the last 3 iterations.

Why the Recall scores were consistently low? Pattern ID 11, ID 34 and ID 36 from Table 4 are not found by the bootstrapping method due to the setting of a window size of 2. Therefore, values extracted by those patterns, which occupy 45% of the total true values, will never be found.

Why many false person-title pairs were included after iteration #8? Domain experts have also designed stop words for each handcrafted patterns, which are capable of screening out common noises with certain patterns. But for the bootstrapping method, those noises extracted by the patterns are still regarded as true.

4.3 Discussions

We find that the bootstrapping method can work only on extracting attribute values of $Title. The values of $Title could be shared by multiple patterns' extractions because multiple people can be assigned to the same position in the government. Only when the values are shared, we can find one pattern with another by bootstrapping. However, one person cannot have multiple fathers and rarely have multiple masters. By now, we have only investigated the pattern-based bootstrapping method in Section 3.2 on the attribute in the task of *attribute discovery*. The preliminary of this method lies in the fact that there should exist some entities that could be extracted by multiple patterns, which makes it possible to find new patterns through pattern generation. However, for the task of *relation extraction* (e.g., *father-son* and *master-disciple*), since each relation pair is unique in the text, there is not a pattern shown in Table 4 that shares even a single common instance that could also be extracted by other patterns in the same category, which makes it hard for instance-level bootstrapping method to work.

# of iterations	P@3	C@3	P@5	C@5	P@7	C@7	P@10	C@10	P@15	C@15	P@20	C@20	AP
1	0.667	0.667	0.800	0.400	0.857	0.429	0.700	0.400	0.467	0.400	0.350	0.300	0.235
2	0.667	0.667	0.800	0.800	0.714	0.714	0.800	0.600	0.667	0.533	0.550	0.500	0.329
3	**1.000**	**0.667**	**0.800**	**0.800**	**0.714**	**0.714**	**0.500**	**0.600**	**0.600**	**0.533**	**0.550**	**0.500**	**0.335**
4	0.667	0.667	0.600	0.800	0.714	0.714	0.500	0.600	0.533	0.533	0.450	0.500	0.279
5	0.667	1.000	0.600	1.000	0.714	0.857	0.700	0.700	0.533	0.667	0.450	0.600	0.320
6	0.333	1.000	0.400	1.000	0.429	0.857	0.500	0.700	0.467	0.667	0.400	0.600	0.254
7	0.333	1.000	0.400	1.000	0.286	0.857	0.500	0.800	0.467	0.733	0.350	0.650	0.214
8	0.333	1.000	0.400	1.000	0.286	0.857	0.500	0.800	0.467	0.733	0.350	0.650	0.204
9	0.000	1.000	0.200	1.000	0.143	0.857	0.400	0.800	0.333	0.733	0.350	0.650	0.162
10	0.000	1.000	0.200	1.000	0.143	0.857	0.200	0.800	0.267	0.733	0.250	0.650	0.131

Table 5: We use *Precision@K*, *Coverage@K* and Average Precision (*AP*) to evaluate the method on pattern extraction. At the iteration #3, the method achieved the highest AP of 0.335, improved relatively 42.6% over the seed iteration (and patterns). However, the reliability of newly extracted patterns significantly reduced and AP started dropping after iteration #5.

5 RELATED WORK

In this section, we survey three main topics related to our work. We point out the uniqueness of our study.

5.1 Chinese NLP Techniques

Though robust NLP techniques are often language independent, most of the NLP techniques for Chinese have their own specific characteristics and thus advantages compared to those for English or other Latin-based languages. Unlike Latin-based languages, Chinese languages do not use white-space as the natural delimiter. Therefore word segmentation is always a key precursor for language processing tasks in Chinese [5, 6, 8, 30, 41, 42]. Moreover, due to lack of morphological features, Chinese Part-of-Speech (POS) tagging and dependency parsing are harder than Latin-based languages like English. Li *et al.* [24] proposed joint models for Chinese POS tagging and dependency parsing tasks. As neural methods have recently achieved significant performance with large amount of annotated data, many deep neural models for Chinese POS tagging and dependency parsing have been developed [9, 21, 33]. Zero Pronoun (ZP) resolution is also a challenging problem in Chinese. Existing studies utilize heuristic rules to resolve ZP issues in Chinese [10, 36]. Recently, supervised neural approaches have been vastly explored on many different tasks [7, 37–39].

However, all these studies focus on modern Chinese text. Classical Chinese is important but was paid little attention, as the majority of precious historical literature was written in classical Chinese hundreds or even thousands of years ago. Doing NLP tasks on classical Chinese would be more difficult than modern Chinese, because of the very different written style and very limited annotated data. Our approach was the first to curate a person entity profiling dataset for the studies and we proposed a pattern-based bootstrapping method to extract the attributes of historical actors in ancient China. The extracted high quality profile information would facilitate history studies. Digital humanities need more attention from both humanity studies and digital technologies.

5.2 Textual Pattern-based Entity Information Extraction Techniques

Given a text corpus, textual patterns leverage statistics (e.g., high frequency) by replacing words, phrases, or entities with symbols such as part-of-speech tags or entity types in order to extract a large collection of tuple-like information [23, 31, 34, 40]. Hearst patterns like "*NP* such as *NP*, *NP*, and *NP*" were proposed to automatically acquire hyponymy relations from text data [14]. Later, machine learning experts designed the Snowball systems to propagate in plain text for numerous relational patterns [1, 4, 43]. Google's Biperpedia [12, 13] generated *E-A* patterns (e.g., "*A* of *E*" and "*E*'s *A*") from users' fact-seeking queries by replacing entity with "*E*" and noun-phrase attribute with "*A*". ReNoun [35] generated *S-A-O* patterns (e.g., "*S*'s *A* is *O*" and "*O*, *A* of *S*,") from human-annotated corpus on a predefined subset of the attribute names. Patty used parsing structures to generate relational patterns with semantic types [29]. The recent MetaPAD generated "meta patterns" based on content quality [16]. However, all the patterns in the above methods can only serve for English. Due to the fundamental grammar difference between classical Chinese and English, the above methods no longer work for our problem. Our work has made the first step in the field of pattern-based entity retrieval that is suitable for classical Chinese text.

5.3 Neural Entity Information Extraction

The task of named entity recognition (NER) is typically cast as a sequence labeling problem and solved by supervised learning models. Different from statistical learning methods like conditional random fields (CRF) [19], end-to-end neural network methods have been proposed to solve the problem [15, 17, 20, 27]. Recent work used language model as another type of supervised signals [25], which can help models obtain more contextual knowledge from corpus without extra annotation. Open source pre-train models have been widely used in the entity information extraction tasks. They improved the performance with models pre-trained on massive corpora. Note that all the models need large amount of annotated data, while unfortunately we don't have in classical Chinese.

6 CONCLUSIONS

In this paper, we aimed at extracting and profiling historical actors from classical Chinese literature. We addressed the challenge of low-resource language. In this study, we employed domain experts to curate a ground-truth dataset of person entities and their

profile attributes and relations (e.g., *courtesy name*, *place of birth*, *title*, *father-son*, *master-disciple*) with handcrafted patterns from two books, *Historical Records* and *Book of Han*. We developed a pattern-based bootstrapping approach to extract the information with a very small number of seed patterns. Experimental results showed the effectiveness and limitations of the iterative method.

ACKNOWLEDGMENTS

The authors would like to thank all the funds for their support. This work was supported in part by Notre Dame Research 2019 Global Gateway Faculty Research Award (RGG) FY19RGG03 373106 and NSF Grant CCF-1901059.

REFERENCES

[1] Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*. ACM, 85–94.
[2] Liang Cai. 2014. Witchcraft and the Rise of the First Confucian Empire. *Albany, NY: State University of New York Press* (2014).
[3] Liang Cai. 2019. Confucians, Social Networks, and Bureaucracy: Donghai Men and Models for Success in the Western Han China (206 BCE–9 CE). *Early China* (2019).
[4] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*.
[5] Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*. Association for Computational Linguistics, 224–232.
[6] Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*. Association for Computational Linguistics, 13–16.
[7] Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In *ACL*. 778–788.
[8] Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of Empirical Methods on Natural Language Processing*. 1197–1206.
[9] Yufei Chen, Sheng Huang, Fang Wang, Junjie Cao, Weiwei Sun, and Xiaojun Wan. 2018. Neural Maximum Subgraph Parsing for Cross-Domain Semantic Dependency Analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. 562–572.
[10] Susan P Converse and Martha Stone Palmer. 2006. *Pronominal anaphora resolution in Chinese*. Citeseer.
[11] Crespigny R. De. 2007. A Biographical Dictionary of Later Han to the Three Kingdoms (23-220 Ad). *Leiden: Brill* (2007).
[12] Rahul Gupta, Alon Halevy, Xuezhi Wang, Steven Euijong Whang, and Fei Wu. 2014. Biperpedia: An ontology for search applications. *VLDB* 7, 7 (2014), 505–516.
[13] Alon Halevy, Natalya Noy, Sunita Sarawagi, Steven Euijong Whang, and Xiao Yu. 2016. Discovering structure in the universe of attribute names. In *WWW*. International World Wide Web Conferences Steering Committee, 939–949.
[14] Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 539–545.
[15] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).
[16] Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M Kaplan, Timothy P Hanratty, and Jiawei Han. 2017. Metapad: Meta pattern discovery from massive text corpora. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 877–886.
[17] Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V Chawla, and Meng Jiang. 2019. The Role of "Condition": A Novel Scientific Knowledge Graph Representation and Construction Model. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1634–1642.
[18] Martin Kern. 2003. The "biography of Sima Xiangru" and the question of the Fu in Sima Qian's Shiji. *Journal of the American Oriental Society* 123, 2 (2003), 303–316.
[19] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).
[20] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*.
[21] Haonan Li, Zhisong Zhang, Yuqi Ju, and Hai Zhao. 2018. Neural character-level dependency parsing for Chinese. In *AAAI*.
[22] Kaiyuan Li. 2000. The Establishment of Han Dynasty and the Liu Bang Group: A Study of the Meritorious Military Class. *Beijing: San lian shu dian* (2000).
[23] Qi Li, Meng Jiang, Xikun Zhang, Meng Qu, Timothy P Hanratty, Jing Gao, and Jiawei Han. 2018. Truepie: Discovering reliable patterns in pattern-based information extraction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1675–1684.
[24] Zhenghua Li, Min Zhang, Wanxiang Che, Ting Liu, Wenliang Chen, and Haizhou Li. 2011. Joint models for Chinese POS tagging and dependency parsing. In *Proceedings of Empirical Methods on Natural Language Processing*. Association for Computational Linguistics, 1180–1191.
[25] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *AAAI*.
[26] Michael Loewe. 2000. A Biographical Dictionary of the Qin, Former Han and Xin Periods: 221 Bc - Ad 24. *Leiden: Brill* (2000).
[27] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016).
[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[29] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: a taxonomy of relational patterns with semantic types. In *Proceedings of Empirical Methods on Natural Language Processing*. Association for Computational Linguistics, 1135–1145.
[30] Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 562.
[31] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of Empirical Methods on Natural Language Processing*. 1499–1509.
[32] Hans Van Ess. 1993. The Meaning of Huang-Lao in Shiji and Hanshu. *Études chinoises* 12, 2 (1993), 161–177.
[33] Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional LSTM. In *ACL*, Vol. 1. 2306–2315.
[34] Xueying Wang, Haiqiao Zhang, Qi Li, Yiyu Shi, and Meng Jiang. 2019. A Novel Unsupervised Approach for Precise Temporal Slot Filling from Incomplete and Noisy Temporal Contexts. In *The World Wide Web Conference*. ACM, 3328–3334.
[35] Mohamed Yahya, Steven Whang, Rahul Gupta, and Alon Halevy. 2014. Renoun: Fact extraction for nominal attributes. In *Proceedings of Empirical Methods on Natural Language Processing*. 325–335.
[36] Ching-Long Yeh and Yi-Chun Chen. 2007. Zero Anaphora Resolution in Chinese with Shallow Parsing. *Journal of Chinese Language and Computing* 17, 1 (2007), 41–56.
[37] Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese Zero Pronoun Resolution with Deep Memory Network. In *EMNLP*. Association for Computational Linguistics, Copenhagen, Denmark, 1309–1318. https://doi.org/10.18653/v1/D17-1135
[38] Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018. Zero Pronoun Resolution with Attention-based Neural Network. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 13–23. https://www.aclweb.org/anthology/C18-1002
[39] Qingyu Yin, Yu Zhang, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2018. Deep Reinforcement Learning for Chinese Zero Pronoun Resolution. In *ACL*. Association for Computational Linguistics, Melbourne, Australia, 569–578. https://doi.org/10.18653/v1/P18-1053
[40] Wenhao Yu, Zongze Li, Qingkai Zeng, and Meng Jiang. 2019. Tablepedia: Automating PDF Table Reading in an Experimental Evidence Exploration and Analytic System. In *The World Wide Web Conference*. ACM, 3615–3619.
[41] Qi Zhang, Xiaoyu Liu, and Jinlan Fu. 2018. Neural networks incorporating dictionaries for chinese word segmentation. In *AAAI*.
[42] Wei Zhou, Aiping Wang, Hua Shu, Reinhold Kliegl, and Ming Yan. 2018. Word segmentation by alternating colors facilitates eye guidance in Chinese reading. *Memory & cognition* 46, 5 (2018), 729–740.
[43] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. 2009. StatSnowball: a statistical approach to extracting entity relationships. In *WWW*. ACM, 101–110.