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

Yihong Ma[1], Qingkai Zeng[2], Tianwen Jiang[2], Liang Cai[3], Meng Jiang[2]

[1]School of Finance, Shanghai University of Finance and Economic

[2]Department of Computer Science and Engineering, University of Notre Dame

[3]Department of History, University of Notre Dame

Motivation

UNIVERSITY OF NOTRE DAME

- Questions of interest by historians:
  - Who came from where?
  - Who studied from whom?
  - Who did what?

- Challenges:
  - Heavy reliance on domain knowledge
    - historians have to spend even longer time learning the specific ancient language first.
  - Time-consuming

Person Entity Extraction

UNIVERSITY OF NOTRE DAME

- Sounds like a simplified NER task?
  - Yes but Not Exactly!
- Different ways of mentioning one specific person:
  - $Hometown + $LastName + $FirstName + $CourtesyName
    - e.g. 蘭陵|孟|喜|長卿; newyork|trump|donald|john
  - $Hometown + $LastName + $FirstName
    - e.g. 蘭陵|孟|喜; newyork|trump|donald
  - $LastName + $FirstName + $CourtesyName
    - e.g. 孟|喜|長卿; trump|donald|john
  - $LastName + $FirstName
    - e.g. 孟|喜; trump|donald
  - $FirstName
    - e.g. 喜; donald

Person Entity Profiling

UNIVERSITY OF NOTRE DAME

- Definition:
  - 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
- Example:
  -

	Meng Xi 孟喜
Courtesy name	長卿
Hometown	東海蘭陵
Title(s)	郎, 曲臺署長, 丞相掾
Father	孟卿
Son(s)	N/A
Master	田王孫
Disciple(s)	趙賓, 白光, 翟牧, 焦延壽

Person Entity Profiling (*cont.*)

UNIVERSITY OF NOTRE DAME

- The great variety of demographic attributes:
  - Each type of attributes needs a set of specific, reliable extractors, which requires prior knowledge of the classical Chinese language.

- Zero Pronoun (ZP):
  - [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.)

  - Not only occur in the same sentence with the mention of the person entity but also across several sentences in the same paragraph.

Key Observation/Assumption

UNIVERSITY OF NOTRE DAME

- 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.

- **Intuition:** In biographical historiography, each chapter is the life story of a certain historical figure.

Handcrafted Patterns

UNIVERSITY OF NOTRE DAME

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

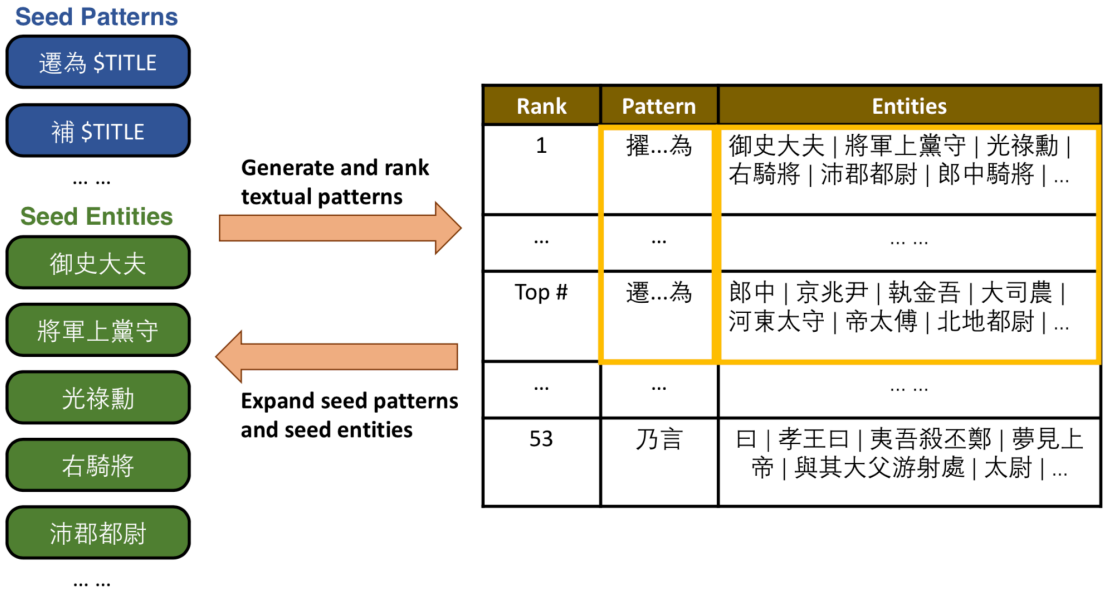
Handcrafted Patterns (*cont.*)

UNIVERSITY OF NOTRE DAME

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

Pattern-based Bootstrapping

- Step 1
  - Generate pattern candidates
- Step 2
  - Rank pattern candidates
- Step 3
  - Select new patterns and extract new values for the next iteration



Generate Pattern Candidates

UNIVERSITY OF NOTRE DAME

- Commonly used *skip-gram* contextual pattern "$w_{-1}\_\_\_w_1$" could hardly work.
  - **Intuition:** target values are more likely to be at the end of the clause because of the linguistic structure.

- We explore two different kinds of contextual features:
  - *$PATTERN $VALUE*
    - A window of a certain size of Chinese characters before a target value.
    - e.g. [遷為 *$TITLE*]
  - *$PATTERN $ENTITY $PATTERN $VALUE*
    - Both a window of one Chinese character before $Entity and all characters between *$ENTITY* and *$VALUE* are selected as the contextual feature.
    - e.g. [遷 *$PERSON* 為 *$TITLE*]

Rank Pattern Candidates

UNIVERSITY OF NOTRE DAME

- Issues when considering all the unlabeled values extracted as false:
  - Mistakenly penalize reliable patterns that extracted true unlabeled values.
  - Could not penalize unreliable patterns that extracted false unlabeled values.
- Ranking function:
  - $w_1 \cdot \frac{\sum_{v \in \mathcal{V}_p} \left(1 - \min_{v^+ \in \mathcal{V}^+} d(v, v^+)\right)}{\sum_{v \in \mathcal{V}_p} freq(v)} + w_2 \cdot \left(1 - \frac{\sum_{v \in \mathcal{V}_p} \max_{v \in \mathcal{V}_p} freq(v)}{\sum_{v \in \mathcal{V}_p} freq(v)}\right) \in [0, 1]$
  - p is a textual pattern; v is a value string; v^+ is a true value string; $\mathcal{V}_p$ is the set of unique strings extracted by pattern p; $\mathcal{V}^+$ is the set of unique true value strings; $d(v_1, v_2)$ is the normalized string distance metric; w_1 and w_2 are weights $s.t. w_1 + w_2 = 1$

Rank Pattern Candidates (*cont.*)

UNIVERSITY OF NOTRE DAME

- $$\frac{\sum_{v \in \mathcal{V}_p} \left(1 - \min_{v^+ \in \mathcal{V}^+} d(v, v^+)\right)}{\sum_{v \in \mathcal{V}_p} freq(v)}$$

  - The textual similarity between the pattern's extracted values and true values

  - **Intuition:** 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.

Rank Pattern Candidates (*cont.*)

UNIVERSITY OF NOTRE DAME

- $1 - \frac{\sum_{v \in \mathcal{V}_p} \max_{v \in \mathcal{V}_p} freq(v)}{\sum_{v \in \mathcal{V}_p} freq(v)}$
  - Variety of the pattern's extracted values
  - **Intuition:** we assume if there is a value whose frequency dominates the set of values that one specific pattern has extracted; the pattern would be less reliable.

Experiment Settings

UNIVERSITY OF NOTRE DAME

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

Evaluating the Handcrafted Patterns

UNIVERSITY OF NOTRE DAME

- Evaluation method:
  - Ground-truth: 15 complete person profiles (with 158 values) annotated by domain experts.
  - Metrics: Precision, Recall, and F1 score.
- Evaluation result:
  - Precision: 0.901; Recall: 0.803; F1: 0.851.
  -

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

Evaluating the Effectiveness of the Bootstrapping Method

UNIVERSITY OF NOTRE DAME

- Evaluating the task of pattern extraction.
- Evaluation the task of person-entity pair extraction.

Evaluating the Task of Pattern Extraction

UNIVERSITY OF NOTRE DAME

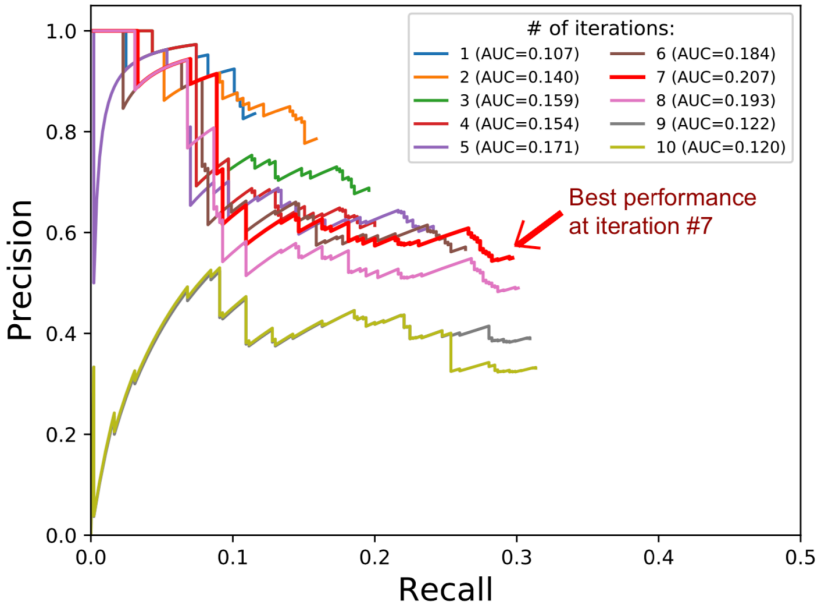
- Evaluation method:
  - Ground-truth: handcrafted patterns for the attribute $*TITLE*
  - Metric:
    - ***P**recision**@K***: the fraction of top K scored generated patterns that are in the ground-truth pattern set.
    - ***C**overage**@K***: the fraction of top K scored ground-truth patterns that are extracted by the bootstrapping method.
    - *Average precision (**AP**)*: the mean of the precision scores after each relevant pattern is retrieved.
- Evaluation result:
  -

# 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

Evaluation the Task of Person-Entity Pair Extraction

UNIVERSITY OF NOTRE DAME

- Evaluation method:
  - Ground-truth: *person-title* pairs extracted by handcrafted patterns.
  - Metric: *precision-recall* curve.
- Evaluation result:
  -



Discussions

UNIVERSITY OF NOTRE DAME

- Bootstrapping only works for *$TITLE*
  - **Preliminary:** there should exist some entities that could be extracted by multiple patterns, which makes it possible to find new patterns through pattern generation.
  - The values of *$TITLE* could be shared by multiple patterns' extractions because multiple people can be assigned to the same position in the government.
  - However, one person cannot have multiple fathers and rarely have multiple masters.
  - Besides, for relation extraction, each relation pair is unique in the text.

Conclusions

UNIVERSITY OF NOTRE DAME

- **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.

Thank You!

UNIVERSITY OF NOTRE DAME

- Any questions?