**Bol — Mining**

<4 figures, 1 table, very tightly set at end>

**Mining and Discovering Biographical Information in *Difangzhi* with a Language-Model-Based Approach**

Bol, P. K., Liu, C.-L. and Wang H.

We present results of expanding the contents of the China Biographical Database[[1]](#footnote-1) by text mining historical local gazetteers, *difangzhi*地方志.[[2]](#footnote-2) The goal of the database is to see how people are connected together, through kinship, social connections, and the places and offices in which they served. The gazetteers are the single most important collection of names and offices covering the Song through Qing periods. Although we begin with local officials, we shall eventually include lists of local examination candidates, people from the locality who served in government, and notable local figures with biographies. The more data we collect, the more connections emerge. The value of doing systematic text mining work is that we can identify relevant connections that are either directly informative or can become useful without deep historical research. Academia Sinica is developing a name database for officials in the central governments of the Ming and Qing dynasties.[[3]](#footnote-3)

# Problem Definition and Main Findings

Figure 1 shows a scanned page from a *difangzhi*,[[4]](#footnote-4) and Figure 2 shows the text of the shown image. As Figure 1 shows, traditional Chinese texts do not have spaces between words or employ punctuation. This feature makes the processing of literary Chinese texts much more difficult than handling alphabetical languages and modern Chinese. The circles in Figure 2 serve as a general delimiter, representing the end of a line, the end of a page, a space, or a transition in text formatting—for example, placing two lines of text in a single column in column 2 of Figure 2.

We would like to algorithmically extract the information about local officials, such as Li Chang 李常in Figure 2. We are interested in the alternative names, such as the style name (*zi* 字) and pen name (*hao*號), birthplace, entry method, serving office, service time, and so forth. In this abstract, we focus on how we identify a person’s name, style name, and the dynasty. For example, we wish to extract the record Song宋as dynasty, Li Chang李常 as person name, and Gongze公擇 as style name after handling the text in Figure 2.

Local gazetteers record various types of information about local areas; we selected those that are related to local government officials. Not counting the circles in the texts, the current study employed 83 text files containing 901,302 Chinese characters.

We extracted 1,260 records from the files and compared them with the biographical data in CBDB. Table 1 gives an analysis of the results, where a circle indicates a match; a cross indicates a mismatch. Among the 1,260 records, 562 match the dynasty, personal name, and style name of some CBDB records, and 544 (43.2%) match only dynasty and name.

# Methods

Figure 3 shows the main procedure for extracting the records. In addition to the gazetteers, we used files of previously known names, addresses, entries, offices, and reign period titles from CBDB to annotate the texts.[[5]](#footnote-5) In this study, we also consider the dynasties for names, offices, and reign periods.

We need to consider the ambiguities of a word when annotating the texts. For instance, ‘Li Chang’ 李常 was a person name in the Song, Yuan, Ming, and Qing dynasties, and the four-character office title ‘Guan-cha tui-guan’ 觀察推官 was an office in the Tang and Song dynasties. In addition the second and third characters ‘cha tui’ 察推also represent an office in the Song and Yuan dynasties. Hence, as illustrated in Figure 4, we could generate at least 16 possible label sequences for the following string T1 in Figure 2.

T1: 李常字公擇南康建昌人自宣州觀察推官發運使

We sift the label sequences by adopting the principle of favoring longer words[[6]](#footnote-6) and by disambiguating with contextual constraints. In T1, we do not consider ‘cha tui’ 察推 an office for the Song dynasty because the four-character sequence is a longer match for the same dynasty. In addition, it is reasonable to require that all labels in a sequence must be *consistent*with the same dynasty. Hence, among the 16 sequences, only ‘李常-觀察推官’ for Song and ‘李常-察推’ for Yuan could survive.

Since there are no known tools for parsing literary Chinese, we employ the concept of language models (Manning and Schütze, 1999) to analyze the texts. We computed, collected, and counted the frequencies of *consistent sequences* of six labels.[[7]](#footnote-7)

Aiming at extracting personal names and style names for government officials, we focused on the consistent sequences that have at least one <NAME> label. We then identified and preferred subsequences that include more different labels. We show four such *filter patterns* below.

P1: <NAME><ADDRESS><REIGN PERIOD><ENTRY>

P2: <NAME><ADDRESS><ENTRY><REIGN PERIOD>

P3: <NAME><NAME><ADDRESS><ADDRESS>

P4: <NAME><ADDRESS><ADDRESS><ADDRESS>

Finally, we selected the consistent sequences that contained the filter patterns and extracted original text segments that corresponded to the consistent sequences. The string T1 was extracted from the text in Figure 2 because it could be annotated with the sequence <NAME><ADDRESS><ADDRESS><ADDRESS><OFFICE><OFFICE>, which contained P4. The <NAME> label is for ‘李常.’ At the annotation stage, our programs did not recognize ‘Gongze’ 公擇as a Style Name for ‘Li Chang’ 李常 because ‘Gongze’ 公擇was not included in the CBDB name list. One of the two annotated results is listed below.

**<NAME Song>李常</NAME>字公擇<ADDRESS>南康</ADDRESS><ADDRESS>建昌</ADDRESS>人自<ADDRESS>宣州</ADDRESS><OFFICE Song>觀察推官</OFFICE><OFFICE Song>發運使</OFFICE>**

To extract ‘Gongze’ 公擇as a Style Name from T1, we parsed the text segment with a low-level grammar pattern for the task. Specifically, a two-character string that appears after the sequence of a <NAME> label and the character ‘字’ (Style Name) and before an <ADDRESS> label was extracted as the Style Name for the <NAME>. With such syntactic rules, we discovered that ‘公擇’ is a Style Name for ‘李常,’ and obtained two records (Song,李常,公擇) and (Yuan, 李常,公擇).

# Results, Evaluation, and Applications

We compared the extracted records with the combinations of dynasty, name, and Style Name in CBDB, and Table 1 shows the results. The two records that we just obtained would belong to type 2, because ‘Gongze’ 公擇 is not known to CBDB. All extracted records of type 2 provide opportunities of finding Style Names that were new to CBDB. However, they should be confirmed by asking a domain expert to check the original text segments, which is an operation facilitated by our software platform.

Extracted records of type 1 do not provide new information if we are just interested in names and Style Names. Certainly, we are more ambitious than this, and type-1 records are instrumental. They help us find the beginnings of the paragraphs that contain extra information about the owners of the type-1 records. T1 is the beginning of the second paragraph in Figure 1. This paragraph contains extra information about ‘李常’ that we can explore to enhance the contents of CBDB. The third paragraph in Figure 1 and many following paragraphs start with statements that we could identify with the filter patterns.

Records of types 3 through 7 make up only about 12.2% of the 1,260 extracted records. Similar to type-2 records, these records do not match any records in CBDB perfectly. After inspecting the original text segments, we will be able to tell whether these mismatches are new discoveries or incorrect extractions.

# Discussion

The reported work represents an extension of our work for CBDB that was reported in Bol et al. (2012). In the previous work, experts manually designed regular expressions for specific text patterns. Now, based on prior information about named entities, we are able to compute and analyze the label sequences for the local gazetteer texts to learn useful filter patterns for automatically extracting desired information. We can apply the reported mechanism to extract birthplaces, service periods, offices, and other basic information, as we just did for extracting names and Style Names. In addition, by identifying key opening statements for paragraphs that contain biographical data, the reported procedure opens a new door for algorithmically extracting information about personal career and social networks. We are working toward learning the document structures of local gazetteers.

Our work is related to automatic grammar induction in computational linguistics. Hwa (1999) learns grammars with data that were manually annotated with syntactic information, and we automatically annotated data with named entities. Klein and Manning (2005) employed advanced techniques to learn hierarchical grammars for Penn treebank sentences, which may be quite challenging in the case of literary Chinese.

Additional responses to reviewers’ comments are available at http://www.cs.nccu.edu.tw/~chaolin/papers/dh2015blw.online.pdf.

# Notes

1. The China Biographical Database (CBDB; http://isites.harvard.edu/icb/icb.do?keyword=k16229) is a collaborative project of Harvard University, Peking University, and Academia Sinica. CBDB is an online relational database with biographical information about approximately 328,000 individuals as of October 2013, primarily from the 7th through the 19th centuries. The data is meant to be useful for statistical, social network, and spatial analysis as well as serving as a kind of biographical reference.

2. *Difangzhi*—see http://www.chinaknowledge.de/Literature/Terms/difangzhi.html.

3. http://archive.ihp.sinica.edu.tw/ttsweb/html\_name/search.php.

4. When written in the vertical style, Chinese paragraphs begin from the right side of a page.

5. Here, ‘names’ include either official names or any alternative names. ‘Addresses’ refer to location names. ‘Entries’ (入仕方式), e.g., ‘進士’ and ‘舉人’, include different ranks and ways of becoming a government official via the Civil Service Examinations (‘科舉’). ‘Offices’ (官職, government positions) include posts in the government. ‘Reign period names’ (年號), e.g., Kangxi 康熙, are names of time periods under a particular emperor.

6. This so-called favoring the longer term 長詞優先principle is commonly adopted when segmenting (or tokenizing) Chinese text strings (cf. Gao et al., 2005).

7. Technically speaking, we are analyzing a 6-gram language model.

# References

**Bol, P. K., Hsiang, J. and Fong, G.** (2012). Prosopographical Databases, Text-Mining, GIS and System Interoperability for Chinese History and Literature. *2012 International Conference on Digital Humanities*, Hamburg, Germany, 19 July 2012.

**Gao, J., Li, M., Wu, A. and Huang C.-N.** (2005). Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics,* **31**(4): 531‒74.

**Hwa, R.** (1999). Supervised Grammar Induction Using Training Data with Limited Constituent Information. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, pp. 73‒79.

**Klein, D. and Manning, C. D.** (2005). Natural Language Grammar Induction with a Generative Constituent-Context Model. *Pattern Recognition,* **38**: 1407‒19.

**Manning, C. D. and Schütze, H.** (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, chap. 6.

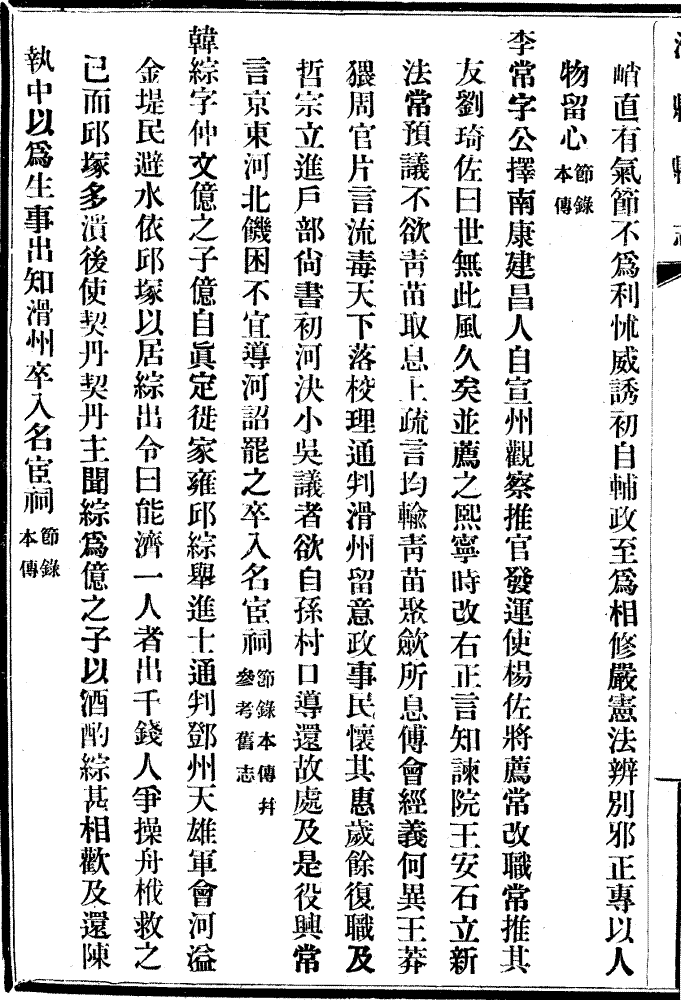


Figure 1.A scanned image of a page of the *Difangzhi* books.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | Dynasty | Name | Style Name | Quan. | Prop. |
| 1 | ○ | ○ | ○ | 562 | 44.6% |
| 2 | ○ | ○ | × | 544 | 43.2% |
| 3 | × | ○ | ○ | 40 | 3.17% |
| 4 | ○ | × | ○ | 31 | 2.46% |
| 5 | × | ○ | × | 29 | 2.30% |
| 6 | × | × | ○ | 20 | 1.59% |
| 7 | ○/× | × | × | 34 | 2.70% |

Table 1. Analysis of the extracted 1,260 records.

○○峭　有氣節不爲利怵威誘初自輔政至爲相修嚴憲法辨别邪正專以人○○○物留心○節錄○本傳○李常字公擇南康建昌人自宣州觀察推官發運使楊佐將薦常改職常推其○○○友劉琦佐曰世無此風久矣並薦之熙寧時改右正言知諫院王安石立新○○○法常預議不欲靑苗取息上疏言均輸靑苗聚歛所息傅會經義何異王莽○○○猥周官片言流毒天下落校理通判滑州留意政事民懷其惠歲餘復職及○○○哲宗立進户部尙書初河決小吳議者欲自孫村口導還故處及是役興常○○○言京東河北饑困不宜導河詔罷之卒入名宦祠○節錄本傳并○參考舊志○韓綜字仲文億之子億自眞定徙家雍邱綜舉進士通判鄧州天雄軍會河溢○○○金堤民避水依邱塚以居綜出令曰能濟一人者出千錢人争操舟栰救之○○○已而邱塚多潰後使契丹契丹主聞綜爲億之子以酒酌綜甚相歡及還陳○○○執中以爲生事出知滑州卒入名宦祠○節錄○本傳○

Figure 2. A passage for the image in Figure 1.



Figure 4. Preferring longer words and context-based disambiguation.



Figure 3. Main steps of the extraction procedure.

1. The China Biographical Database (CBDB; http://isites.harvard.edu/icb/icb.do?keyword=k16229) is a collaborative project of Harvard University, Peking University, and Academia Sinica. CBDB is an online relational database with biographical information about approximately 328,000 individuals as of October 2013, primarily from the 7th through the 19th centuries. The data is meant to be useful for statistical, social network, and spatial analysis as well as serving as a kind of biographical reference. [↑](#footnote-ref-1)
2. *Difangzhi*—see http://www.chinaknowledge.de/Literature/Terms/difangzhi.html. [↑](#footnote-ref-2)
3. http://archive.ihp.sinica.edu.tw/ttsweb/html\_name/search.php. [↑](#footnote-ref-3)
4. When written in the vertical style, Chinese paragraphs begin from the right side of a page. [↑](#footnote-ref-4)
5. Here, ‘names’ include either official names or any alternative names. ‘Addresses’ refer to location names. ‘Entries’ (入仕方式), e.g., ‘進士’ and ‘舉人’, include different ranks and ways of becoming a government official via the Civil Service Examinations (‘科舉’). ‘Offices’ (官職, government positions) include posts in the government. ‘Reign period names’ (年號), e.g., Kangxi 康熙, are names of time periods under a particular emperor. [↑](#footnote-ref-5)
6. This so-called favoring the longer term 長詞優先principle is commonly adopted when segmenting (or tokenizing) Chinese text strings (cf. Gao et al., 2005) [↑](#footnote-ref-6)
7. Technically speaking, we are analyzing a 6-gram language model. [↑](#footnote-ref-7)