Toward Algorithmic Discovery of Biographical Information in Local Gazetteers of Ancient China

Chao-Lin Liu[†] **Chih-Kai Huang**[§] **Hongsu Wang**[‡] **Peter K. Bol**[!]
[†§]Department of Computer Science, National Chengchi University, Taiwan
[‡!]Institute for Quantitative Social Science, Harvard University, USA
[†]Graduate Institute of Linguistics, National Chengchi University, Taiwan
{[†]chaolin,[§]102753029}@nccu.edu.tw, {[‡]hongsuwang, [!]pkbol}@fas.harvard.edu

Abstract[1]

Difangzhi (地方志) is a large collection of local gazetteers complied by local governments of China, and the documents provide invaluable information about the host locality. This paper reports the current status of using natural language processing and text mining methods to identify biographical information of government officers so that we can add the information into the China Biographical Database (CBDB), which is hosted by Harvard University. Information offered by CBDB is instrumental for human historians, and serves as a core foundation for automatic tagging systems, like MARKUS of the Leiden University. Mining texts in *Difangzhi* is not easy partially because there is litter knowledge about the grammars of literary Chinese so far. We employed techniques of language modeling and conditional random fields to find person and location names and their relationships. The methods were evaluated with realistic *Difangzhi* data of more than 2 million Chinese characters written in literary Chinese. Experimental results indicate that useful information was discovered from the current dataset.

1 Introduction

Person and location names are two crucial ingredients for studying historical documents. Knowing the participants and locations provides a solid foundation for detecting and reasoning about the developments of historical events. Detecting temporal markers is also very important for historical studies, yet, for Chinese history, it is relatively easier to spot the temporal markers because the names of the dynasties and reign periods (年號, nian2 hao4) are known and stable.

We apply techniques of natural language processing and machine learning to find person names, location names, and their relationships in *Difangzhi* (**地方志**, **DFZ** henceforth) in the present work, aiming to enrich the contents of the China Biographical Database (Bol, 2012). DFZ is a general name for a large number of local gazetteers that were compiled by local governments of different levels in China since as early as the 6th century AD (cf. Hargett, 1996). DFZ contain a wide range of information about their host locations, and the biographical information about the government officers is our current focus.

The main barrier for achieving our goals is that there is little completed work in the literature about the grammars for literary Chinese, while grammars are central for extracting named entities like person and location names from texts with computational methods (Gao et al., 2005; Nadeau and Sekine, 2007).

Figure 1 shows the image of a sample DFZ page. In the old days, Chinese texts were written from top to bottom and from right to left on a page. Most linguists know that there are no word boundaries in modern Chinese. It might be quite surprising for researchers outside of the Chinese community that there were even no punctuations in literary Chinese. Without clear delimiters between words and sentences, it is very challenging even for people to read literary Chinese, so it takes a serious research to find ways not just for segmenting words but also for splitting sentences in literary Chinese (Huang et al., 2010).

Grammar induction (de la Higuera, 2005) is a general name for enabling computers to learn the grammars of natural languages. Some researchers worked on the grammars for selected sources of Chinese. Huang et al. (2001) ex-

[1] An extended version of this paper appears in the proceedings of the Third Big Humanities Data Workshop in the 2015 IEEE Int'l Conf. on Big Data (Liu et al., 2015). The main contents of this paper and the workshop paper are the same, while the workshop paper is an extended version.

29th Pacific Asia Conference on Language, Information and Computation pages 87 - 95
Shanghai, China, October 30 - November 1, 2015

Figure 1. A page of DFZ

plored the induction problem with about a thousand sentences that were extracted from Hanfeizi (韓非子) and Xunzi (荀子), both of which are classics that are more than two thousand years old. Kuo (2009) tried to find phrase-structure rules for modern Chinese texts, and Lee and Kong (2012) built treebanks for Tang poems. Although these researchers worked on grammars for Chinese texts, they encountered Chinese patterns that are quite different from the ones that we need to handle in DFZ.

Previous works for inducing grammars of literary Chinese employed some forms of pre-existing information to begin the induction procedures. Given that literary Chinese texts consisted of just long sequences of characters, the needs for external information for grammar induction should be expected. Hwa (1999) assumed that the training corpus was partially annotated with high-level syntactic labels. Lü et al. (2002) started with bilingual corpora. Yu et al. (2010) embarked with a sample treebank, and Boonkwan and Steedman (2011) began with some syntactic prototypes.

We tackle the NER tasks in literary Chinese from two unexplored perspectives. First, we employ the biographical information in the China Biographical Database (**CBDB**, henceforth) to annotate the DFZ texts, learn language models (**LMs**, henceforth) from the annotated texts, and extract biographical information based on the learned models. Alternatively, we train conditional-random-field (Sutton and McCallum, 2011) models with a set of labeled DFZ data that were achieved in (Bol et al., 2012; Pang et al., 2014), and use the conditional-random-field (**CRF**, henceforth) models to extract candidate names from the test data, which is another set of DFZ texts. We have verified the findings of the LM-based and the CRF-based methods. Both show very good results for NER in DFZ.

We present the sources of our data, define our target problems, and discuss the motivation for our work in Section 2. We then provide details about our main approaches in two long sections. In Sections 3 and 4, we look into details about the LM-based and CRF-based methods, respectively, including the designs of the classification models and results of several evaluation tasks. In Section 5, we wrap up this paper with a brief summary and discussions about some technical issues.

2 Data Sources, Problem Definitions, and Motivation

We provide information about the sources of our data, define the problems that we wish to solve, and explain the rationality of our approaches in this section.

2.1 Unlabeled Data

Currently, we have two sets of DFZ text files. The unlabeled part has more than 900 thousand of characters that were extracted from 83 volumes of local gazetteers (Bol et al., 2015). The labeled part will be presented in Section 2.4.

These 83 volumes were compiled between the middle of the Ming dynasty (1368-1644AD) and the early Min-Guo period (since 1912AD). These books were produced by governments of different levels at 65 locations in China.

Figure 1 shows a sample page from this collection. It is hard to count the number of columns on this page. Typically, we consider that a column, in this case, consists of two thinner columns. A person name is emphasized by occupying the width of a column, and details about this person are recorded in the thin columns. Therefore, we would say that the leftmost three columns of text in Figure 1 would read like the passage shown in Figure 2.

The DFZ texts may contain characters that are not or seldom used in modern Chinese. If these characters have modern counterparts, they will be substituted by their modern replacements; otherwise, spaces will take their positions. As an example for the former case, the eleventh character on the second column from the right in

不知勞洪武元年楊璟取廣西吉尼堅壁不下城破執送京師不屈死郡人感其德立廟祀之陳瑜字仲庸雷州人廣西中書省都事城破以佩刀自刎有劉永錫者潭州人與瑜同事率妻子溺於白龍池死焉曾尚賓江西人為義兵千户洪武元年明兵圍靜江尚賓守西城城陷身中數鎗知不敵

Figure 2. A partial DFZ passage from Figure 1

Figure 1 is "裏" (li3), which may be written as "裡" (li3) in our files. When the latter cases occur, understanding the original DFZ records will become even more challenging.

2.2 Problem Definitions

We wish to build a system that can extract biographical information from DFZ to enrich the contents of CBDB. The current contents of CBDB were extracted from sources other than DFZ (Pang et al., 2014). Hence, we are interested in spotting all types biographical information in DFZ.

In this paper, we focus on issues about finding person names and location names, and extend to some relevant topics, such as checking whether the locations were native places. In the longer run, we will expand our attention to find information about social networks and personal careers as well.

2.3 More on Motivation

For a text passage as illustrated in Figure 2, it is very challenging for people to find useful information without assistive information, even for modern generations of native speakers of Chinese. In the text file, it is not easy to find the name "陳瑜" (chen2 yu2) that was written in larger characters in the original DFZ.

The grammars of literary and modern Chinese are not exactly the same, and reading literary Chinese is a lot harder than reading modern Chinese, especially when there are no boundary markers between sentences. In addition, historical knowledge is also required for correct word segmentation and lexical disambiguation, which are important for understanding and extracting desired information from the texts.

To achieve our goals, we need some informative sources for the work of information extraction. The importance of these informative sources for our methods for extracting information is just like that of the machine-readable dictionaries for the methods for handling modern natural languages.

Our approaches are innovative because we utilize the biographical information in CBDB to provide semantic information about the DFZ texts. In contrast, the literature that we reviewed in Section 1 carried out grammar induction with such linguistic knowledge as part-of-speech tags and syntactic structures.

2.4 Labeled Data

We have a set of labeled DFZ data. This set of data was collected from 143 volumes of DFZ, which contained more than 1498 thousand of characters.

The DFZ texts were labeled with regular expressions (**REs**, henceforth) that were compiled by domain experts (Bol et al., 2012; Pang et al. 2014), and the REs were designed to extract biographical information. The labeled data were then saved as records in a large table with 113,784 records in total.

Each record has many fields, and the fields were designed to contain a wide variety of factoids about the individuals. Major fields contain information about an individual's legal name, style name (字, zi4), pen name (號, hao4), dynasty, native place (籍貫, ji2 guan4), serving office (官職, guan1 zhi2), entry method (入仕方法, ru4 shi4 fang1 fa3), service time, service location, and reign period (年號, nian2 hao4).

Due to the nature of the original DFZ data and the limited expressiveness of REs, a non-negligible portion of the fields do not have values (i.e., have missing values), and, sometimes, the values are not correct. Nevertheless, these labeled data remain to be valuable and prove to be useful from the perspectives of historical studies (Pang et al., 2014) and of building machine-learning models.

3 Language-Model-based Approach

We annotate the DFZ with the biographical information available in CBDB, and find the frequent and *consistent* *n*-grams for locating candidate strings from which we may extract legal names and style names.

3.1 Labeling and Disambiguation

Figure 3 lists the steps of our main procedure, **Constrained N-Grams** (**CNGRAM**), for NER. First, we label the text with information in CBDB. Five types of labels are in use now: **name** for a legal or a style name, **address** for locations, **entry** for entry methods, **office** for

Procedure CNGRAM (txt, idbs, cc)
txt: DFZ texts
idbs: information databases
cc: chosen conditions for checking consistency
Steps
1. Label the text based on the given ***idbs***. Prefer the labels that cover longer strings, all else being equal.
2. For contexts of chosen conditions, ***cc***, remove the inconsistent labels.
3. Find the frequent consistent *n*-gram patterns, and use them as filter patterns
4. Try to extract named entities from strings that conform to the filter patterns

Figure 3. The CNGRAM procedure

service office, and **nianhao** for reign periods.

In reality, some strings may be labeled in more than one ways. For instance, "陽朔" (yang1 shuo4) can be a reign period of the Han dynasty or a location name, and "王臣" (wang2 chen2) is a very popular person name that was used in many dynasties. Before we try to disambiguate the labeling, we will keep all possible labels for a string.

We will use the following short excerpt of Figure 2 to explain the execution of CNGRAM.

T1: 陳瑜字仲庸雷州人廣西中書省都事

Identifying T1 from its context is possible because this string begins and ends with words that have corresponding labels. We will find out that there was a person named "陳瑜" (chen2 yu2) in Yuan, Ming, and Qing dynasties and that both "雷州" (lei2 zhou1) and "廣西" (guang3 xi1) were addresses.

At the first step of CNGRAM, we prefer longer matches for the same type of labels, as a heuristic principle for disambiguation. The principle of preferring longer words is very common for Chinese word segmentation. In T1, both "中書省都事" (zhong1 shu1 sheng3 dou1 shi4) and "中書省" can be labeled as office names in the Yuan dynasty, but we would choose the former because "中書省都事" is a longer string. In contrast, we do not have "中書省都事" for the Ming dynasty, so will use "中書" and "都事" for Ming.

We also assume that named entities in a passage should be ***consistent*** in some senses, as another heuristic principle for disambiguation. This consistent principle should be reminiscent of the "**one sense per discourse**" principle for word sense disambiguation (Yarowsky, 1995).

Currently, we presume that named entities in a context of six labels should be referring to something of the same dynasty, where six is an arbitrary choice and can be varied. We have not used addresses to check consistency because we are still expanding our list of addresses. Therefore, we do not accept a "陳瑜" of the Qing dynasty because neither "中書省都事" nor "中書" is an entity in Qing. Using the consistent principle, we will keep labels only for the Song and Ming dynasties for the sample passage, thereby achieving some disambiguation effects.

Hence, we have two consistent sequences: [name("陳瑜", Yuan), address("雷州"), address ("廣西"), office("中書省都事", Yuan)] and [name("陳瑜", Ming), address("雷州"), address("廣西"), office("中書", Ming), office("都事", Ming)].

3.2 Extracting Unknown Style Names

Aiming at extracting person and style names for government officers, we focus on the consistent sequences that have at least one **name** label. We then identify and prefer strings that are associated with more different labels. We show four such ***filter patterns*** below.

P1: name-address-nianhao-entry
P2: name-address-entry-nianhao
P3: name-name-address-address
P4: name-address-address-office

These patterns shed light on how person names were presented in DFZ texts. We can now examine the DFZ strings that are labeled with these patterns to judge whether these patterns indeed carry useful information. Usually, we find regularities in these statements, and can implement specific programs for extracting target information from such patterns.

Our running example, T1, contains the P4 pattern in two different ways, and we list the substrings.

T2: 陳瑜字仲庸雷州人廣西中書
T3: 陳瑜字仲庸雷州人廣西中書省都事

In both T2 and T3, we see that a key signal "字" (zi4), which is a typical prefix for style names, follows a **name** label. "字" is followed by two unlabeled characters which are then followed by an **address**, an unlabeled character, another **address**, and an **office**. Thus, T2 and T3 are examples of pattern P5, where <name> and <address> represent labeled strings and Z1 and Z2 are two unlabeled characters.

P5: <name> 字 Z1 Z2 <address>

The unlabeled characters, Z1 and Z2, can be extracted as style names because practical statistics indicate that over 98% of style names contain exactly two characters. Therefore, we embody this finding with actions in our programs.

The third step in CNGRAM is thus an interactive step[2], and requires human participation. Notice that the work for domain experts is quite minimal and that the results are worthwhile. A human expert does not have to read 83 books to find the candidate patterns. Using CNGRAM to locate string patterns that contain useful information, we are able process a large amount of data both efficiently and effectively.

With the extracted style name "仲庸" (zhong4 yong1), we can create two records, i.e., (Yuan, 陳瑜, 仲庸) and (Ming, 陳瑜, 仲庸). "仲庸" is unknown to CBDB, and can be added to CBDB with the approval of domain experts.

The CNGRAM procedure actually helps us learn the language models that were used in DFZ. By inspecting frequent and consistent patterns that actually contain biographical information, we can gather more knowledge about grammar rules in DFZ and then implement NER procedures based on the observations.

3.3 Empirical Evaluations

We compared the extracted records with the records in CBDB (2014 version) to evaluate the CNGRAM procedure, and show the results in Table 1, where the circles and crosses, respectively, indicate matches and mismatches between the extracted and CBDB records.

The matching results are categorized into types, e.g., type 1 is the group that we had perfect matches in dynasty, legal name, and style name. We have 609 such instances in the current experiment, and the proportion of type-1 instances is 28.3% of the 2152 extracted records.

The two records that we obtained in the previous subsection belong to type 2, because "仲庸" is not known to CBDB. All extracted records of type 2 provide opportunities of finding unknown style names for CBDB. However, they should be confirmed by historians. The experts may check the original texts for this approval procedure, which is an operation facilitated by our software platform.

Table 1. Analysis of 2152 extracted records

Type	Dynasty	Name	Style N.	Quan.	Prop.
1	○	○	○	609	28.30%
2	○	○	×	665	30.90%
3	×	○	○	117	5.44%
4	○	×	○	262	12.17%
5	×	○	×	220	10.22%
6	×	×	○	45	2.09%
7	○/×	×	×	234	10.87%

Records of types 3 and 4 are similar to records of type 2. They offer opportunities of adding extra information to CBDB. Records of types 5, 6, and 7 provide some opportunities for adding information about new persons in CBDB. After inspecting the original text segments, we will be able to tell whether these mismatches are new discoveries or just incorrect extractions.

3.4 Further Extensions

We are more ambitious than verifying whether CNGRAM can help us find correct biographical information. Type-1 records can be instrumental for advanced applications. They help us find the beginnings of the descriptions that contain information about the owners of type-1 records.

If we can determine the beginnings of two consecutive segments, then we can find persons who have relationships. T1 is the beginning of a major segment in Figure 1. The string "也兒吉尼字尚文唐兀氏人" is the beginning of a segment for a person named "也兒吉尼" (ye3 er2 ji2 ni2). The person mentioned between "也兒吉尼字尚文唐兀氏" and T1, e.g., "楊璟" (yang2 jing3), should have some relationships with "也兒吉尼".

In addition, it is quite intriguing to apply pattern P5, in Section 3.2, in an extreme way. Figure 4 shows the raw data for the text in Figure 2. If we compare Figure 4 and the image in Figure 1 carefully, we can find that the circles were added to signify (1) changes between major columns and thin columns or (2) changes of lines. The semantics of the circles are ambiguous, but they are potentially useful.

If "字" is really a strong indicator that connects legal names and style names, P6 and P7 may lead us to find pairs of legal and style names. Here, we use C1, C2, and C3 to denote Chinese characters.

P6: ○ C1 C2 C3 字 Z1 Z2
P7: ○ C1 C2 字 Z1 Z2

[2] Using computers to select the patterns is possible if we are willingly to set a frequency threshold to determine "frequent" patterns, which may not be a perfect choice for historical studies.

◯不知勞洪武元年楊璟取廣西吉尼堅壁不下◯城破執送京師不屈死郡人感其德立廟祀之◯陳瑜◯字仲庸雷州人廣西中書省都事城破以佩刀自刎◯有劉永錫者潭州人與瑜同事率妻子溺於白龍池◯死◯焉◯曾尚賓◯江西人為義兵千户洪武元年明兵圍靜◯江尚賓守西城城陷身中數鎗知不敵自◯

Figure 4. A partial DFZ passage with circles

When we find substrings that conform to P6 or P7 in the raw data, we may want to check whether C1-C2-C3 (or C1-C2) is a legal name, Z1-Z2 is a style name, and their combination is for a real person.

We evaluated this heuristic approach with the unlabeled data of Section 2.1, and found 3765 pairs of (legal_name, style_name). We checked these pairs with CBDB (2014 version), and achieved Table 2. Note that strings conforming to P6 and P7 have very short contexts, so we could not judge the dynasties of these names, so Table 2 is simpler than Table 1.

Table 2 shows that 31% of the pairs have corresponding records in CBDB. Although we cannot guarantee the correctness of these matched records, the statistics are promising and encouraging. 1192 type-1 records matched the legal and style names of certain CBDB records. This amount is more than the number of type-1 records in Table 1. Some of the pairs that we identify with the current heuristic did not appear in filter patterns that we discussed in Section 3.2, suggesting that a hybrid approach might be worthy of trying in the future.

4 CRF-based Approach

CRF-based models (Sutton and McCallum, 2011) are very common for handling NER with machine learning methods (Nadeau and Sekine, 2007). We employed MALLET (McCallum, 2002) tools for building linear-chain CRF models, and trained and tested our models with the labeled data that we discussed in Section 2.4.

4.1 Features

Given the training data (cf. Section 2.4) and the biographical information in CBDB, we can create a feature set for each Chinese character in DFZ for training and testing a CRF model. We consider four types of features: original characters, relative positions of named entities in CBDB, whether the character was used in person or location names in DFZ, and whether the characters belong to a named entity.

Table 2. Analysis of 3765 extracted records

Type	Name	Style Name	Quan.	Prop.
1	◯	◯	1192	31.66%
2	◯	×	885	23.51%
3	×	◯	1104	29.32%
4	×	×	584	15.51%

We explain our features listed in Table 3, using T3, in Section 3.2, as a running example.

The original Chinese characters are basic features, summarized in groups 1 and 2 in Table 3. For the position of "州" (zhou1), "州" is an obvious feature for itself. The surrounding ***k*** characters can be included in the feature set as well. If we set ***k*** to three, the three characters before and after "州", i.e., "仲" (zhong4), "庸" (yong1), "雷" (lei2), "人" (ren2), "廣" (guang3), and "西"(si1), are included in the feature set.

Relative positions of the closest named entities (**NEs**) are summarized in group 3 in Table 3. We consider four types of NEs: office names, entry methods, reign periods, and time markers, and will record NEs on both sides of the current position. The first three types are just like the office, entry, and nianhao labels that we discussed in Section 3.1. The time markers refer to a special way of counting years in China, i.e., Chinese sexagenary cycle (干支, gan1 zhi1), and names of months when they were used. We consider NEs that are within 30 characters on either side of the current position, so a position can have up to eight features of group 3.

In T3, there are three characters between "州" and "中書省都事", so **officeRight@3** would be used as a feature for "州". The label name consists of three parts: the type of NEs, the direction respective to the current position (i.e., Right or Left), and the number of characters between the current position and the NE.

Group 4 is about the usage of the current position. It would be helpful to know the probability of the current character being used in a person name or in a location name. Equation (1) shows the basic formula.

$$\Pr(x \text{ in person names}) = \frac{freq(x \text{ in person names})}{freq(x \text{ in } DFZ)} \quad (1)$$

In T3, "雷" is used as a character in a location name. We calculated the frequency of "雷" being used in location names, and divided this frequency by the total frequency of "雷" in DFZ. We discretized the probability measure into five equal ranges: [0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6,

Table 3. Features for CRF models

Group	Types	Description
1	Chinese characters	self
2	Chinese characters	surrounding ***k*** characters
3	relative positions of selected named entities	office, entry, reign period, and time
4	usage	used in person or location name
5	usage	family name?
6	named entities	office, entry, reign period, and time

0.8), and [0.8, 1.0]. If the probability of "雷" was used in a location name was 0.45, we would add probLoc@3 for "雷", where 3 means the third interval in the discretized ranges.

Group 5 is also about the usage of the current position. There is a list of well-known Chinese family names, that is commonly called Hundred Family Names (百家姓, bai3 jia1 xing4)[3]. We add a feature to the current position if it is in the list. In T3, "陳" (chen2) is such a character. If a family name has two characters, the features will indicate the positions of the characters, e.g., "歐" (ou1) and "陽" (yang2) in "歐陽" will, respectively, have surename@1 and surename@2 as their features.

Features in group 6 are for four types of the named entities, i.e., **office**, **entry**, **nianhao** (for reign period), and **time** (as we discussed for the features in group 3). In general, historians have more complete information about these key types of NEs in Chinese history, so using specific tags for these NEs may offer stronger signals for nearby person and location names.

When we used group 6 along with other groups, we would not annotate a position with features in groups 1 through 5, if the position is part of certain named entities of group 6. Instead, we would use only the values for group 6. For example, at the beginning of the text in Figure 2, we have "洪武元年楊璟" (hong2 wu3 yuan2 nian2 yang2 jing3), where "洪武" represented a reign period, so both characters would be annotated only by **nianhao**. "楊璟" did not belong to any types of NEs in group 6, so would be annotated with other features.

Features of groups 3 and 6 are related in nature. We will see that using group 6 in places of group 3 led to better performance in the next section.

4.2 Evaluation: Labeled Data

We evaluated the effectiveness of using linear-chain CRF models for recognizing person and location names in DFZ with the labeled data that was discussed in Section 2.4. Given the original labels, we could create feature sets for all characters, and then ran 5-fold cross validations.

We classified the characters into seven categories: NB, NI, NE, AB, AI, AE, and O. We use N and A to denote name and location, respectively. B, I, and E denote beginning, internal, and ending, respectively. O means others. Hence, for example, NB is for the first character of a person name and AE is the last character of a location name.

We ran several experiments for CRF models that considered different combinations of the features that we discussed in Section 4.1. The classification results were measured by standard metrics, i.e., precision, recall, and F_1 measure that are very common for information retrieval.

Table 4 shows the experimental results for four such combinations. The results improved gradually for the experiments listed from the left to the right side. The first row of Table 4 lists the combinations of features used in the experiments. The second row shows the abbreviated names of the performance measures. The leftmost column shows the seven categories of the classification results.

The experiments that used only group 1 as the feature provided results that were better than we thought. Identifying categories of individual characters in the dataset of Section 2.4 did not seem to be a very challenging task. We added the second group of features by setting ***k*** to five. Then, we added group 4, group 5, and group 6, one by one for the listed experiments.

We did add group 3 in some of our experiments, but adding this group generally made the experimental results worse than not having them, so we do not show those results.

We also set ***k*** to three and seven, but we did not observe significant differences in the results. Setting ***k*** to seven provided a bit better results, but the improvement was not statistically significant.

Recognizing the categories of individual

[3] http://baike.baidu.com/subview/6559/15189786.htm

Table 4. Performances of selected CRF models

	Group 1			Groups 1+2			Groups 1+2+4+5			Groups 1+2+4+5+6		
	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1
O	0.96	0.90	0.93	0.97	0.94	0.95	0.97	0.96	0.96	0.97	0.97	0.97
NB	0.76	0.91	0.83	0.85	0.94	0.89	0.91	0.94	0.93	0.93	0.95	0.94
NI	0.78	0.85	0.82	0.86	0.91	0.88	0.91	0.92	0.91	0.93	0.93	0.93
NE	0.72	0.87	0.79	0.82	0.92	0.87	0.89	0.92	0.90	0.91	0.93	0.92
AB	0.78	0.83	0.80	0.85	0.86	0.86	0.89	0.88	0.88	0.91	0.89	0.90
AI	0.48	0.73	0.57	0.71	0.84	0.77	0.80	0.86	0.83	0.83	0.89	0.86
AE	0.79	0.83	0.81	0.85	0.86	0.86	0.89	0.88	0.88	0.91	0.89	0.90

characters was just a basic task for our system. Our goal was to identify person names and location names. Hence, we really care about whether a sequence of NB, NI, and NE, for instance, indeed represented a person name.

We conducted such an integrated verification with the best performing model in Table 4, i.e., using groups 1, 2, 4, 5, and 6 as features. A name, either for a person or for a location, must exactly match the original labels, to be considered as a correct classification. For person names, the precision and recall rates are 92.0% and 93.9%, respectively. For location names, the precision and recall rates are 91.0% and 89.5%, respectively. Fining location names is harder than finding person names.

4.3 Evaluation: Unlabeled Data

We trained a CRF model, employing feature groups 1, 2, 4, 5, and 6, with all of the labeled data (Section 2.4), and evaluated the model with the task of identifying person and location names in the unlabeled data (Section 2.1). Due to the page limit, we cannot report the results.

5 Discussions and Concluding Remarks

We reported our work for mining biographical information from *Difangzhi* with techniques of language models and conditional random fields. Results observed in practical evaluations proved the effectiveness of these technologies for named entities recognition in literary Chinese.

As illustrated in Figures 1 and 2, processing texts of literary Chinese with computer programs is challenging. We approach this problem with gradually more complex methods. Building our current work on the data that were labeled in previous work (Bol, 2012; Pang et al., 2014) and CBDB, we were able to apply LM and CRF based models. The CNGRAM (Section 3.1) is an interactive procedure that was designed to guide researchers to find useful patterns.

For practical applications, the LM and CRF models may be integrated with an online tagging service, MARKUS (Ho, 2015)[4], of the Leiden University. As we collect more information about person names, style names, pen names, location names, and native places, we become more competent to separate the continuous Chinese strings into meaningful paragraphs (cf. Section 3.4) and find social networks of the government officers (Bol et al., 2015).

In the near future, we will consider more domain-dependent knowledge and contextual constraints to recognize and disambiguate named entities. People of different dynasties may have the same name, for instance. In a *Difangzhi* book, records about government officers of the same dynasty usually appeared close to each other. Many times, the records were ordered chronically. Considering these constrains for disambiguation can make our annotations about a person more precise.

In the longer run, mining the grammar rules of literary Chinese is a bigger and rewarding challenge. It was found that the language models and CRF models worked better for some of the 83 *Difangzhi* books but not as well for others (Bol et al., 2015). People who compiled these books adopted different styles of writing, and the styles varied from time to time and from place to place. Knowing the grammar rules that govern these language patterns will enable us to find more precise information from *Difangzhi* and perhaps other historical documents written in literary Chinese.

Acknowledgements

This work was supported in part by the Ministry of Science and Technology of Taiwan under grants MOST-102-2420-H-004-054-MY2 and MOST-104-2221-E-004-005-MY3. We thank the reviewers for their valuable comments, with which we can strengthen our work in the future.

[4] http://chinese-empires.eu/analysis/tools/

References

Bol, Peter K. 2012. Historical research in a digital environment, keynote speech in the 3rd International Conference on Digital Archives and Digital Humanities, <http://isites.harvard.edu/fs/docs/icb.topic1080143.files/Historical%20Research%20in%20Dig%20Env.ppt>.

Bol, Peter K., Jieh Hsiang, and Grace Fong. 2012. Prosopographical databases, text-mining, GIS and system interoperability for Chinese history and literature, *Proceedings of the 2012 International Conference on Digital Humanities*.

Bol, Peter K., Chao-Lin Liu, and Hongsu Wang. 2015. Mining and discovering biographical information in *Difangzhi* with a language-model-based approach, Presented in the 2015 International Conference on Digital Humanities.

Boonkwan, Prachya and Mark Steedman. 2011. Grammar induction from text using small syntactic prototypes, *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 438–446.

de la Higuera, Colin. 2005. A bibliographical study of grammatical inference, *Pattern Recognition*, 38: 1332–1348.

Gao, Jianfeng, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach, *Computational Linguistics*, 31(4): 531–574.

Hargett, James M. 1996. Song dynasty local gazetteers and their place in the history of Difangzhi writing, *Harvard Journal of Asiatic Studies*, 56(2):405–442.

Ho, Hou Ieong. 2015. MARKUS: A fundamental semi-automatic markup platform for classical Chinese, Presented in the 2015 International Conference on Digital Humanities.

Huang, Hen-Hsen, Chuen-Tsai Sun, and Hsin-Hsi Chen. 2010. Classical Chinese sentence segmentation, *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 15–22.

Huang, Liang, Yinan Peng, Huan Wang, and Zhenyu Wu. 2001. PCFG parsing for restricted classical Chinese texts, *Proceedings of the 1st SIGHAN Workshop on Chinese Language processing*, 1–6.

Hwa, Rebecca. 1999. Supervised grammar induction using training data with limited constituent information, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 73–79.

Kuo, Yu-Chen. 2009. *Using Reinforcement Learning to Learn Phrase Structure Parsing in Mandarin Chinese*, Master's Thesis, National Tsing Hua University, Taiwan. (in Chinese)

Lee, John and Yin Hei Kong. 2012. A dependency treebank of classical Chinese poems, *Proceedings of 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 191–199.

Lü, Yajuan, Sheng Li, Tiejun Zhao, and Muyun Yang. 2002. Learning Chinese bracketing knowledge based on a bilingual language model, *Proceedings of the 19th International Conference on Computational Linguistics*, 1–7.

Liu, Chao-Lin, Chih-Kai Huang, Hongsu Wang, and Peter K. Bol. 2015. Mining local gazetteers of literary Chinese with CRF and pattern based methods for biographical information in Chinese history, *Proceedings of the 3rd Big Humanities Data Workshop* in 2015 IEEE International Conference on Big Data, accepted.

McCallum, Andrew Kachites. 2002. MALLET: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>

Nadeau, David and Satoshi Sekine. 2007. A survey of named entity recognition and classification, *Linguisticae Investigationes*, 30(1):3–26.

Pang, Wai-him, Shih-pei Chen, and Hui Cheng. 2014. From text to data: Extracting posting data from Chinese local monographs, *Proceedings of the 5th International Conference on Digital Archives and Digital Humanities*, 93–116.

Sutton, Charles and Andrew McCallum. 2011. An introduction to conditional random fields, *Foundations and Trends in Machine Learning*, 4(4):267–373.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189–196.

Yu, Kun, Yusuke Miyao, Xiangli Wang, Takuya Matsuzaki, and Junichi Tsujii. 2010. Semi-automatically developing Chinese HPSG grammar from the Penn Chinese Treebank for deep parsing, *Proceedings of the 23rd International Conference on Computational Linguistics*: posters, 1417–1425.