Automatic Biographical Information Extraction from Local Gazetteers with Bi-LSTM-CRF Model and BERT

Zhou Liu[a], Hongsu Wang[b], and Peter K. Bol[c]
[a]Department of Philosophy, Peking University, pkulz@pku.edu.cn [b]Institute for Quantitative Social Science at Harvard, Harvard University, hongsuwang@fas.harvard.edu
[c]Department of East Asian Languages and Civilizations, Harvard University, peter_bol@harvard.edu

Abstract

Named entity information in Chinese local gazetteers supports extending the Chinese Biographical Database (CBDB) project. Instead of using regular expressions method and manual work to tag biographical information using LoGaRT, we propose an automatic deep learning method that uses tagged data to train a Bi-LSTM-CRF model that can then be applied to an untagged dataset without manual work. This method can not only dramatically improve tagging efficiency, but also overcome the shortcomings of regular expression in named entity justification by utilizing semantic information. Moreover, we employ the advanced pre-trained language model, BERT, to encode our word vectors and further improve performance. This method has performed very well on our local gazetteers dataset and extracted data for CBDB. This experiment can also support our further work on unstructured, narrative historical data and demonstrates the applicability of deep learning methods to ancient Chinese texts.

Keywords: Local Gazetteers, NER, BERT, Bi-LSTM-CRF

1 Introduction

CBDB (Chinese Biographical Database) is a relational database that stores biographical information about approximately 470,000 individuals who lived primarily from the 7th to the 19th centuries in China. Not merely serving as a dictionary of biographical information, CBDB users can search for a specific individual's date of birth and death, biographical address, kinship and social associations, career information, and more, as

well as build more indirect and complex queries to obtain information about people based on area, time, status, and relationship. Then, such data can be used to conduct prosopographical, social network, and spatial analysis. Support for such comprehensive quantitative research necessitates the continual expansion of the database through an ongoing data collection program.

One of the key pre-tasks for database construction is to choose reliable data sources. Major data sources for the CBDB project include both semi-structured and unstructured texts, such as biographical dictionaries and indexes, the local histories known as "local gazetteers," occasional writing from individual literary collections, epitaphs, and formal dynastic sources (Fuller, 2020). Among these data sources, we preferentially choose texts with a relatively clear structure, to which regular expressions can be applied and large amounts of data can be effectively extracted (Chen, 2013). Apart from regular expression, we also explored other data mining methods, such as an image pre-processing algorithm (Tsui, 2017) and an LSTM network (Long short term memory) (Han et al., 2019).

Though the regular expression method worked well for our data mining tasks, exploration of the use of deep learning methods to extract data from ancient Chinese historical texts have begun to reveal their potential as the technology continues to develop. Rapidly developing NLP (natural language processing) models can "learn" the semantics of text very well. In this sense, by introducing deep learning methods, we can utilize texts' meaning rather than merely their syntax and text structures for data mining. We'll discuss the advantage of semantic information for NER (named entity recognition) in more detail with the example of local gazetteers, which is the main motivation of our experiment and paper.

One of our new data sources is a collection of 3017 digitized local gazetteers from East View, from which we initially collect records of local officials from 600 BCE to 1900 BCE. Local gazetteers are comprehensive records kept in multiple volumes that include data on the natural environment (landscape, meteorology, hydrology, crops, etc.), local government institutions and offices held, religious sites, schools and students, civil service degree holders, and biographies. Our research focused on the section devoted to local officials (职官志), which records the names of persons holding official positions in a given locality under different dynasties and often includes further information on their places of origin, titles, dates of service, and more.

To extract this data, the CBDB project first collected the digital texts of 1800 local gazetteers. Initially, we developed an interactive platform, SmartRegex, for users to tag and output data (Pang et al., 2018). SmartRegex allows both customized regular expressions to tag all strings of text that conform to a particular pattern and the manual fine-tuning of each individual tag. While using SmartRegex, we were required to first separate out the text strings pertaining to each individual by using regular expression to recognize a person's surname, which in Chinese texts appears before the given name. We then summarized the structural pattern of records in that gazetteer or set of gazetteers and wrote a custom regular expression to tag the information that shares the same pattern in batch. When the information recorded did not strictly follow the same pattern, we had to once again tag each item manually. After the information was tagged, it was output as spreadsheets by SmartRegex. Using this procedure, it took 3 years to process 1800 local gazetteers with 241,148 local official records. However, there are still 3017 digitized local gazetteers that require processing, as well as many more existing local gazetteers to be digitized and studied.

The development of LoGaRT is based on SmartRegex (see also the paper on LoGaRT by Chen in this issue). Tagging with LoGaRT offers substantial advantages, but here we highlight some shortcomings and propose some ways to deal with these problems.

The first problem is that this method cannot be as efficient when applied to a large amount of data. The more text data we have, the less uniformity there is. Thus, more regular expressions and manual work are required to complete the tagging task.

The second problem is directly related to the disuse of semantic information. A regular expression searches for a matched string with a specific pattern, stipulating the beginning and end, the length, and such formal features. This method fails when there is no distinct pattern for name entities. For instance the sentence 李甲北直隸人 (Li Jia, who was born in North Zhidi) contains a person name and his place of birth. However, there is no specific partition between these two entities; instead, the name and the town location are simply stated. Thus, regular expression cannot be sure whether 李甲 is a person name and 北直隸 is a place, or 李甲北 is a person name and 直隸 is a place.

Another complication with using regular expression is that it cannot handle a full series of tags at one time, as the occurrence order of different tags is not uniform in local gazetteers. For instance, sometimes information about place appears after a person's

name, while sometimes information about time appears after a person´s name, and we cannot give a certain regular expression pattern to handle these two situations at once. This diversity also increases with the increasing amount of local gazetteer data.

In essence, the heart of the problems with using regular expression lies in the fact that regular expression can only recognize formal features; it cannot "understand" the meaning of a named entity. To avoid this shortcoming, we turned to a deep learning method to let the neural network "learn" the tagging rules by itself instead of giving a specific tagging pattern. A neural network model can take the tagged text as input, give a predicted tag and compare it with the actual tag, then update the "rules" that are represented by parameters and apply it to its prediction in the next iteration. This is one round of training. With hundreds of rounds of training, the trained model gains its own "understanding" of the tagging rules and can tag new text automatically without additional knowledge.

Compared with regular expression, the process of training and learning does not require any invention of specific rules supplied by professional knowledge. Instead, it learns by itself with the aim of minimizing the loss between predicted result and actual tags. What the model learns is not only the specific signals of some patterns, but also semantic relationships. For example, the neural network can learn that some characters are very likely to represent a place, while others are just function words, and the probability of the order of occurrence of different tags (e.g. place usually appears after name). Another advantage of the deep learning method is that it can learn the rules of all tags, and tag all the tags at one time, rather than tagging each tag separately.

In this paper, we made an experiment with deep learning methods on local gazetteers from which we have extracted using the regular expression method. We employed the Bi-LSTM (Bidirectional Long Short Term Memory)-CRF (Conditional Random Field) model to complete this tagging task. The training data for this was derived from a CBDB sub-project led by Chen Shih-Pei, Chen Hui, and William Pang. These data were tagged using regular expressions and proof-read by scholars (Pang et al., 2018). To ensure precision in tagging, we utilized a state-of-the-art pre-trained language model, BERT (Bidirectional encoder representations from transformers), a state-of-art language model, to provide the word embedding.

As for the paper´s structure, we will first introduce the related studies on local gazetteers and NER (Named-entity recognition) tasks on classical Chinese texts. Second,

we will explain our datasets and specific tagging tasks, and then we will describe our model architecture, training process, and evaluation. The code and training models are available at https://github.com/zhouliupku/LGtagging_LSTM. Our paper aims at explaining the data processing and model training, and proving that this deep learning method effectively works on local gazetteers. In addition, this experiment can support further exploration of data mining in narrative biographical texts.

2 Related Work

As mentioned above, LoGaRT is a state-of-the-art platform for tagging Chinese local gazetteers in order to support quantitative historical studies that use data extracted from local gazetteer texts (Chen et al., 2017). It relies on scholarly judgment to identify the differences in the presentation of the same type of information within records and expressions but promises to extract all the sought-after data nearly automatically thereafter. LoGaRT aims at providing a humanist-friendly platform with which scholars can themselves tag the information they seek and output the data for further quantitative analysis. However, with increasing demand for diverse data in large quantities, the degree of human supervision necessary for this process makes this platform inefficient.

From the deep learning perspective, this task of information extraction from local gazetteers can be naturally formulated as a Name Entity Recognition (NER) task. NER systems have been studied and developed widely for decades. Among them, the Bi-LSTM-CRF model (Huang et al., 2015) is one of the most advanced works. Thus, we choose to apply this model to our task in this paper.

This model is effective and versatile. It combines the use of bi-direction to memorize both past and future input features, as well as a Conditional Random Field (CRF) layer to use sentence level tag information, thus dramatically improving sequence tagging performance. For traditional Chinese texts, this model can be applied to sentence segmentation, word segmentation, and parts of speech tagging tasks; it performs well even on cross-dynastic datasets (Cheng et al., 2020).

Another technique that boosts performance in NER tasks is word representation. Recent work in this area has explored embeddings that carry complicated semantic information about words, yet are easy to manage, among which are pre-trained word representations like ELMo (Embeddings from language models) (Peters et al., 2018), GPT (Generative Pre-Training) (Radford et al., 2018) and BERT (Devlin et al., 2019).

The most recent and state-of-the-art model, BERT uses Transformer (Vaswani et al., 2017) for encoding to improve computing efficiency beyond ELMo. It also fully utilizes bi-directional information through a masked language model rather than sequence prediction as in GPT. Moreover, BERT provides a pretrained model for various languages and convenient interfaces for downstream tasks like sentence pair classification, question answering, and tagging.

As for NER tasks in traditional Chinese texts, various models have been applied in previous research to different genres of texts. The Conditional Random Fields (CRFs) model has been applied to Chinese classic novels from the Ming and Qing dynasties and the model achieved an F-score in a range of 67% to 80% among the novels (Long et al., 2016). Long's work only concentrated on classic novels in order to cover more text characteristics. Yan's work (Yan et al., 2020) generated their own dataset with tagging tools from different genres and proposed a CNN-based model, MoGCN (Mixture of Gated Convolutional Neural Network), under the assumption that a local block based upon n-gram character bundles would show a stronger capacity to capture explicit semantic information than the word-level feature extraction used to process the NER task. Their method showed a 1.5% F-score improvement over other models. From these studies we can conclude that the deep learning method can be effectively used on traditional Chinese texts, while optimization in NER models, especially models performing better in semantic generation, can improve tagging accuracy. Thus, transformers and BERT application to traditional Chinese texts is a new tendency and focus. BERT has already been applied to encode traditional Chinese text for sentence segmentation and punctuation tasks. In Yu's paper (Yu et al., 2019), sentence segmentation was essentially formulated as a tagging task with a small tag set containing [B] for sentence beginning and [I] for in the sentence. A similar framework was used for the punctuation task, which is of great significance for traditional texts in which both manuscript and print did not include punctuation. The results of this experiment also indicate that the pre-trained BERT for Modern Chinese texts can be extended to traditional Chinese texts. Although pioneering in utilizing tagging networks in traditional Chinese contexts, it focused only on sentence level tasks without considering more sophisticated tagging at the entity level. Our model proves the effectiveness of BERT both at a sequence level and an entity level.

3 Dataset and Task Specification

In this section, we will introduce our training dataset and the two main tasks we focused on in this study. The first task was to separate the whole page into sentences/records, and the second was to tag the name entities within each record.

3.1 Basic Definitions and Task Specification

Information on local officials from a digitized local gazetteer appears in LoGaRT as seen below:

○郭○治○江西泰和舉人嘉靖二年任○陞嵩明知州見名宦○○董○傑○南直贛榆人○監生嘉靖七年任○○王○校○江西泰和舉人嘉靖九年○任清苦自持卒於官○○鄧文憲○廣東新會舉人嘉靖十一年以○御史謫任陞淮安同知○○胡○崗○南直蕪湖人監○生嘉靖十三年任○○傅○錠○江西進賢舉○人嘉靖十四年任○○陰○積○福建監生○嘉靖十五年任○○孫○本○通州舉人○嘉靖十七年任○○陸○經○號西湖南直華亭○舉人嘉靖二十二年任○○黃延年○廣東南海舉人○嘉靖二十三年任○○○○○○○○○○○○○井一成○建德人監生嘉靖二十四年任慷○慨能詩文調靖安縣知縣○○朱敬之○四川峽江舉人嘉靖二十六年任清○廉慈惠調廣東茂名縣知縣○○林○植○廣東東莞舉人嘉靖二十九○年任陞東平州知州○[i]

LoGaRT notes three levels, of which the "Record" is the first and basic level, followed by "Page" and "Section":

- Record: a record is a sequence of text starting with a named entity. In the example above, the first record in the body of text above is "○郭○治○江西泰和舉人嘉靖二年任○陞嵩明知州見名宦○ ". The "○" in the original text represents a blank space in the original undigitized text. This is simply a blank; it is not a white space marking word segmentation, and it does not bear any semantic information. Each record contains similar semantic meanings: a person named NAME(○郭○治 = ○ Guo ○Zhi), born in some PLACE(江西泰和 = Jiangxi[province] Taihe [county]), who entered government service through the ENTRY TYPE(舉人 = juren, someone who passed the provincial civil service examination) at a specific TIME(嘉靖二年 = Jiajing [reign period] 2nd year [i.e.,1523]), the official position he was assigned after this current official position (陞嵩明知州 = promoted to magistrate of Songming subprefecture) and some additional information that does not appear in this example, such as other names, kinship and social associations, and so on. Though all records share similar

contexts, the specific expressions of each record are not exactly the same in Chinese. Thus, we cannot just rely on unified regular expression to extract all the information we want.

- Page: a page is a sequence of records (as in the text above) under a unique page ID. There is no clear separation (punctuation and paragraph) between each record within the same page in the original text.
- Section: a section is a sequence of pages stored in the same document. It can also be regarded as a chapter.

With LoGaRT, the text is tagged like this (Figure 1), in which colors identify different semantic units:

○郭○治○江西泰和舉人嘉靖二年任○陞嵩明知州見名宦○
○董○傑○南直贛榆人○監生嘉靖七年任○
○王○校○江西泰和舉人嘉靖九年○任淸苦自持卒於官○
○鄧文憲○廣東新會舉人嘉靖十一年以○御史謫任陞淮安同知○
○胡○崗○南直蕪湖人監○生嘉靖十三年任○
○傅○錠○江西進賢舉○人嘉靖十四年任○○
陰○積○福建監生○嘉靖十五年任○
○孫○本○通州舉人○嘉靖十七年任○
○陸○經○號西湖南直華亭○舉人嘉靖二十二年任○
○黃延年○廣東南海舉人○嘉靖二十三年任○○○○○○○○○○○○○
井一成○建德人監生嘉靖二十四年任慷○慨能詩文調靖安縣知縣○
○朱敬之○四川峽江舉人嘉靖二十六年任淸○廉慈惠調廣東茂名縣知縣○
○林○植○廣東東莞舉人嘉靖二十九○年任陞東平州知州○

Figure 1

LoGaRT uses colors to represent tag types, which can be customized by the user. In this example, blue represents the person's name, red represents a biographical address, light blue represents the entry type, pink represents posting time, and so on. The remaining text, shown in black, does not contain any of the desired information. To

complete this task automatically, we must first separate sentences at the beginning of a person's name and then tag each entity we need.

3.2 Dataset size

We randomly split the dataset into training, cross-validation, and test sets, with a ratio of 6:2:2.

Table 1

	Training set	CV set	Test set
Page Count	16682	5560	5562
Record Count	151336	50445	50447
Character Count	2801090	929055	925796

3.3 Data Pre-processing

The "○" in original text represents a blank space in non-digitized text; it does not bear any semantic information. Because we wanted the model to learn semantic information without relying on any hints about patterns, such as given by specific symbols like "○", we first removed all such symbols in the digitized text.

As preparation for training the model, we first added tags to each character. For page model, i.e. separating page into record, we added "S" for the last character of the record, and "N" for other characters. For record model, i.e. extracting name entities in the record, we parsed tagged data into BIO-style tags, in which "B" represented the beginning of a named entity, "I" represented the middle of the named entity, and "O" meant outside of the named entity. Following this, for each input sequence, we also added special indicators for the beginning [CLS] and end [SEP] of a sentence, with corresponding tags to satisfy the input format of BERT.

In cases where padding was needed, for example in batch training that required the same length of records in the same matrix, we added placeholders with [PAD] tags. Lastly, for Unicode characters outside the range of Chinese characters (0x4e00 to 0x9FFF), we substituted [UNK].

4 Model

4.1 Model Architecture

We used Bi-LSTM-CRF as the architecture for the neural network. On top of that, before the LSTM layers, we use the BERT model to embed sentences with length *N* into an embedded input of size *N*×*M*, where M = 768 is the output dimension of BERT.

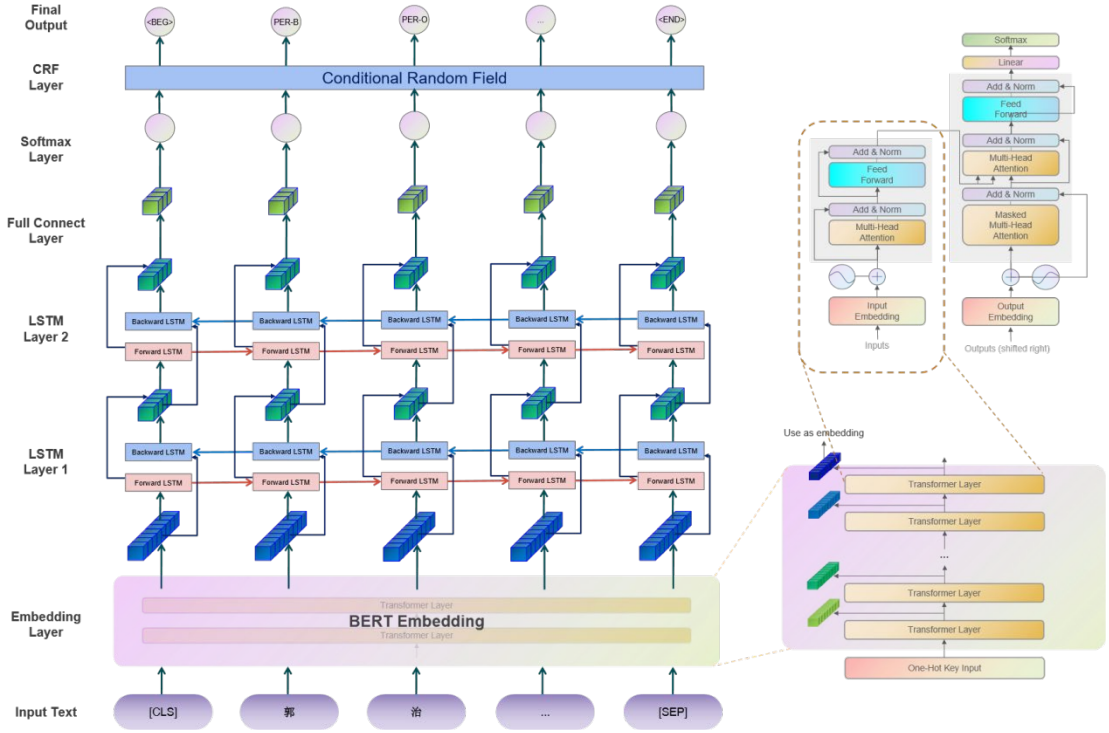

Figure 2

Figure2 above is an illustration of the model. The first layer is an embedding layer, which can turn the Chinese characters into vectors. In the NLP area, word embedding represents words with a high level of semantic correlation using word vectors that are closer in the vector space. Take our daily words as an example. The distance between the vector of the word “apple” and the vector of “pear” is much smaller than the vector distance between “pear” and “bear” because “apple” and “pear” are both fruits but “pear” and “bear” are different categories even though they are similar in form. As technology continues to advance, more pre-trained language models have produced promising performances in multiple classic NLP tasks. These pre-trained models can not only save users time on word embedding training, but also show great reliability in specific tasks based on the huge volume and diversity of their training corpus. In our study, we used

BERT because it employed the state-of-art training models and techniques and has shown excellent performance in various NLP tasks.

The next layer is a bidirectional LSTM layer. An LSTM network can intake the input in sequence, then predict the status at t_n based on its "memory" of previous information and the input x_n at t_n. For instance, it can predict "治" with a name tag based on the word vector of "治" and the previous information that "郭" is also a name tag. The bidirectional LSTM takes the input sequence forward and backward, so that it can utilize the memory from both previous information and later information. For instance, it can predict "治" should have a name tag based on the word vector of "治", the previous information that "郭" is also a name tag, and the back information that "江" is a place tag, "西" is a place tag, and so on. Here we established two bidirectional LSTM layers to enhance the memory through the course of our research.

The full connection layer is for dimension reduction, from the LSTM layer dimension to the dimension of number of tags, so that the softmax layer can take in such information and give the probability distribution on all tags. Figure 3 below is a full connection illustration:

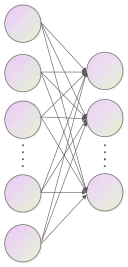

Figure 3

Figure 4 illustrates the process of softmax. The softmax layer maps the output from the full connection layer to a probability distribution within (0, 1). For instance, in the chart below, the softmax layer can map the vector (-2.6, -1.8, 0.1, -1.9, 0.5) to vector (0.66, 0.04, 0.04, 0.00, 0.25). In the mapped vector, each number represents the probability of each tag, and the tag with the maximum probability is the prediction.

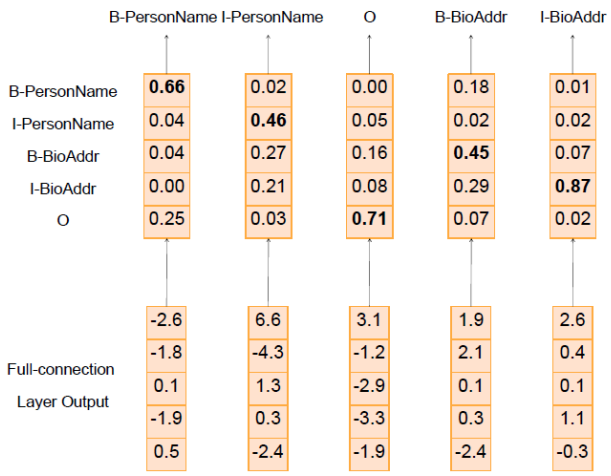

Figure 4

After all terms above, we added a CRF layer that can attach some constraints to the final predicted tags to ensure their validity. For instance, in the CRF layer, the model can learn that the tag of the first word in a sentence should start with "B-" or "O", not "I-"; in a "B-label$_1$ I-label$_2$ I-label$_3$ I-..." sequence, label$_1$, label$_2$, label$_3$ should be the same named entity tag, and any other information based on the whole sentence level. These constraints can be learned from the CRF layer automatically based on the training dataset during the training process.

4.2 Two Step Tagging

Due to the text structure of our task, we streamlined the tagging as two steps: page model and record model. Page model refers to the network that handles the text separation task by splitting a page into records. The aim of the record model was then to tag different name entities within a record. The tags that we used were {B, I} × C, where *C* is all the classes including person name, other alternative names, courtesy name, style name, biographical address, kinship, entry time, entry address, entry type, posting time, office title, posting type, previous office title, and next office title.

Models for the two steps were trained separately. For practical use, we would apply the two models in one after the other to get fully tagged results.

5 Experiments and Performance

5.1 Evaluation Metrics

To measure the performance of text separation, the prediction of end-of-sentence flag was the sole concern. Therefore, we measured the precision (P), recall (R), and F-score (F) with respect to this flag.

To measure the performance of the tagging task, we focused on the entity level measures, i.e. for entity type *i*, TP_i indicated the number of exact matches to objects with this entity type (i.e. correct beginning and ending location, correct entity type), FP_i indicates the number of spurious objects with this entity type, and FN_i indicates the number of missing objects with this entity type. As a hypothetical example, if the true labels were "[B-PersonName] 郭 [I-PersonName] 治 [B-BioAddr] 江 [I-BioAddr] 西 [I-BioAddr]泰[I-BioAddr]和[O]人"and our model gave the prediction as "[B-PersonName] 郭[I-PersonName]治[B-BioAddr]江[I-BioAddr]西[O]泰[O]和[B-PersonName] 人", we evaluated it as having one TP entity, one FN entity, and one FP entity. In this example, true labels form two entities, "郭治" and "江西泰和", the first of which was correctly predicted (hence counted as one TP) but not the second (hence one FN) even though there was a partial match. Moreover, "人" was not actually a part of any entity used here, but prediction makes it an extra entity, and therefore it was counted as one FP.

Note that the tagging task in the record model was a multi-class classification task, so aggregation was needed in order to give an overall assessment. Two main frameworks have been widely used (Sokolova et al., 2009). Macro averaging considers the average of the same measures for each class, while micro averaging calculates the cumulative *TP*, *FP*, and *FN* by summing up the counts of each in all classes, therefore putting heavier weight on the more frequently occurring classes. In our research, we adopted micro averaging, because for downstream research relying on the output, it is the preferred method to obtain accurate information about commonly appearing fields, such as person name and biographical address. Specifically, for a total of N_T entity types, precision (P), recall (R), and F-score (F) were defined as following:

$$\mathrm{P}_{\mu} = \frac{\sum_{i=1}^{N_T} TP_i}{\sum_{i=1}^{N_T} TP_i + FP_i}$$

$$R_\mu = \frac{\sum_{i=1}^{N_T} TP_i}{\sum_{i=1}^{N_T} TP_i + FN_i}$$

$$F_\mu = \frac{2}{{P_\mu}^{-1} + {R_\mu}^{-1}}$$

5.2 Page Model Performance

We trained the aforementioned neural network with various choices of hyperparameters, and used the combination with the best performance on the cross validation set (learning rate = 3×10^{-6}, LSTM dimension = 64, LSTM layer = 2, batch size = 4). Performance per epoch is illustrated in Figure 5 below.

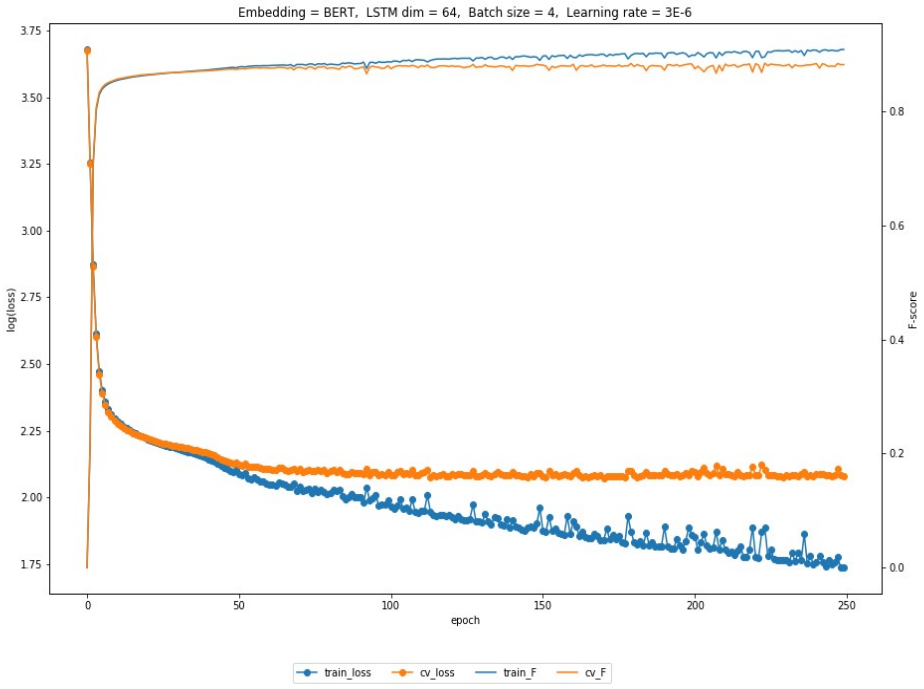

Figure 5

Both F-score (right axis) and loss (left axis) on the cross validation converged relatively fast with highest F-score achieved on epoch 223. Table 2 shows evaluation on the training, cross validation, and test sets.

Table 2

Metric	Train	CV	Test
Precision	91.85%	90.22%	89.13%
Recall	87.38%	86.76%	86.06%
F-score	89.56%	88.46%	87.57%

5.3 Record Model Performance

Using a methodology similar to the page model, we trained the record model with multiple combinations of hyperparameters, and illustrated the performance of the best model (learning rate = 10^{-6}, LSTM dimension = 64, LSTM layer = 2, batch size = 4) in Figure 6.

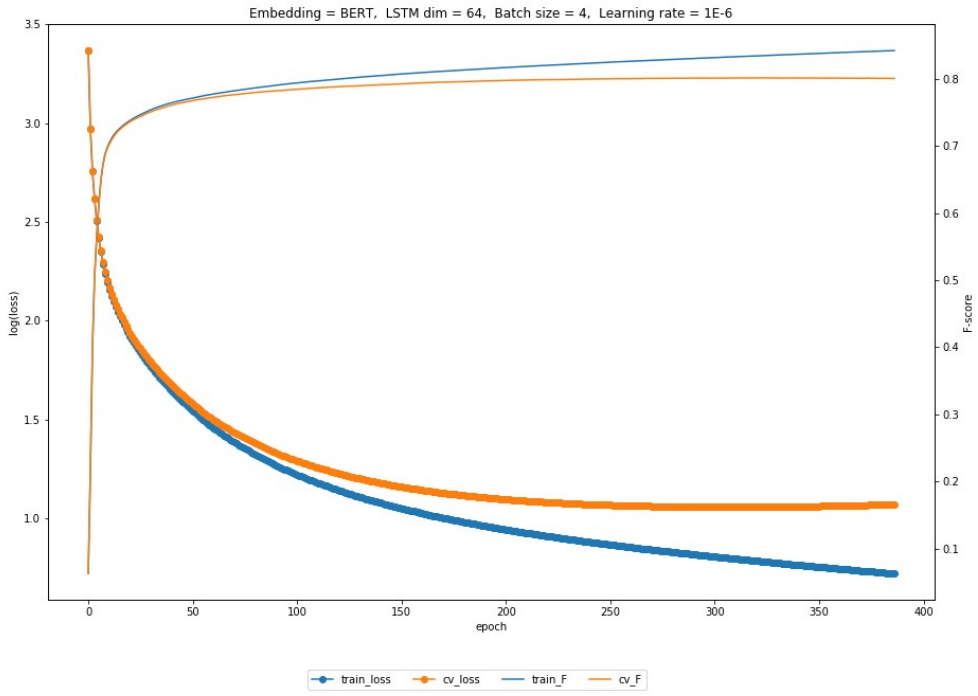

Figure 6

Here, convergence was slightly slower than in the page model but still within reasonable iterations. Table 3 shows the performance on different subsets.

Table 3

Metric	Train	CV	Test
Precision	84.86%	81.64%	81.38%
Recall	82.15%	78.75%	78.78%
F-score	83.49%	80.17%	80.06%

5.4 Tag Improvement analysis

Table 4

Tag	Post time	Person	Jiguan(birth place)	Other name	Dy(dynasty)
#Real tag	37722	50447	31431	173	28
#Correctly Predicted	32098	41850	21854	48	3
#Correct Ratio	85.09%	82.96%	69.53%	27.75%	10.71%

In our experiment, there are up to 18 categories of entities, which also increase the difficulty of learning and tagging. Though we achieved an overall F-score over 80%, results among different tags varied greatly and led to instructive conclusions. Table 4 above shows some representative tagging results on test data.

The "post time" tag shows the top correct ratio. The reason can be analyzed from two main aspects. For one thing, "post time" frequently appears in records and thus has relatively more training data. Also, the expressions of "post time" in local gazetteers are limited, and thus easy to "learn."

Compared with the "post time" tag, the "person" tag has an even greater frequency in the data but shows a lower correct ratio due to the complexity of person name expressions. However, it still performs well with an over 80% correct ratio, which proves that our model can learn that the "meaning" of such characters is a name, rather than only distinguish the structure of a name, because the name expressions are too diverse to generalize an uniformed expression structure.

Further evidence can be deduced from the "jiguan" tag. At the beginning of our paper, we gave an example sentence, "李甲北直隸人" (Li Jia, who was born in North Zhidi Provence). Our model can tag "李甲"(Li Jia) as a person name and "北直隸"(North Zhidi Provence) as "jiguan"(birthplace). Such sentence structure, i.e. a birthplace appearing after person name without any separate symbols is a common

expression in local gazetteers. The correct ratio of the “jiguan” tag is nearly 70% and it can prove that our model can “understand” the distinction between person name and birthplace based on semantic information. In this sense, the deep learning model can process a very common case that regular expression cannot handle.

Finally, we noticed that “other name” tag and “dynasty” tag gain a very low correct ratio due to the lack of sufficient training data.

5.5 Application

With the trained model, we set up a demo (https://github.com/zhouliupku/LGTaggingApp) to present the tagging results intuitively to make it easy for readers to use (note that this tool is for Chinese historical local gazetteers only).

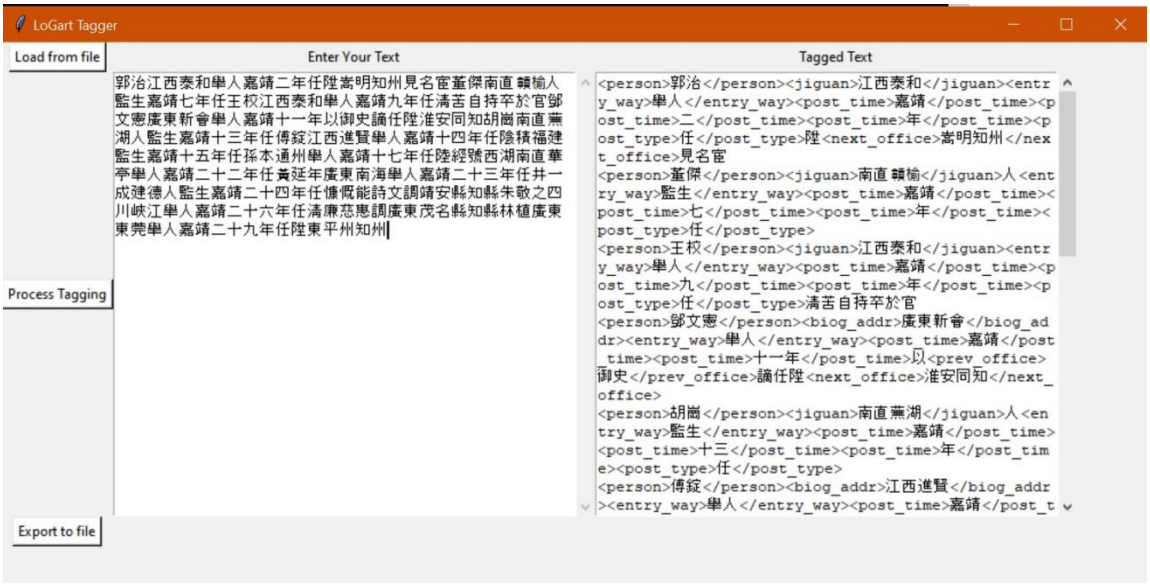

Figure 7

As shown in the illustration in Figure 7, this program supports both loading a text from a file and entering the text, as well as presenting the result and exporting the results to a file. The input is the plain text without any Chinese padding characters, punctuation, or extra formatting. The outputs are in a form that can be put directly into LoGaRT and presented with tags and colors. Also, this program supports exporting the file in Excel form for further analysis.

6 Conclusion and Future works

To conclude, by contrast to the regular expression method, which processes named entity tagging tasks from the expression structure level, the deep learning method advances these tasks to the semantic level. Thus, with Bi-LSTM-CRF and BERT we trained a supportive system for tagging data in local gazetteers. This system can be reused with different local gazetteer texts and greatly improves the efficiency of entity tagging when applied to large number of texts. With this auto-tagging system we can obtain reliable results and greatly reduce the need for manual checking.

There are several interesting ways to enhance the potential of our current model. First of all, we used classical settings to separate the original data, i.e., 60% of data was used as the training set. This is a significant amount: the number of records was as large as 151K. Given the amount of manual work that was required to collect, tag, and review the training data, a natural question is whether it is possible to train a comparable model with a smaller training dataset while still maintaining robustness. Research on the relationship between performance scores like F-value and training data size may help in deciding how much data must be manually tagged in training for NER tasks. To search for a smaller data size that can maintain accuracy, we could employ the method of bisection on training data.

Another feature of the local gazetteer dataset that we were not able to explore deeply is its internal structure within and among dynasties and places. Records from the same period, for example within the same dynasty, share similarities both in style and in syntax, as do those pertaining to the same place. In this paper we have tried to train a one-size-fits-all model that can handle the whole dataset. However, to better utilize the dynastic and geographic structure, it would be interesting to see if performance could be improved by first clustering records and then training sub-models for each category.

Finally, we believe this method can be applied not only to more newly digitized local gazetteers, but also to other biographical texts that cannot be fully processed by regular expression because of unclear expression structures.

Bibliography

Chen, S. (2013). Text extraction using regular expressions. In *Digitization in the Humanities Workshop*. Rice University.

Chen, S.-P., Che, Q., Ling, C., Schäfer, D., and Wang, H. (2017). Treating a genre as a database: the Chinese local gazetteers, the LG tools, and research based on this new digital methodology. In Lewis, R., Raynor, C., Forest, D., Sinatra, M., and Sinclair, S., editors, *Digital Humanities 2017: conference abstracts*, pages 53–54. McGill University & Université de Montréal, Montréal, Canada.

Cheng, N., Li, B., Xiao, L., Xu, C., Ge, S., Hao, X., and Feng, M. (2020). Integration of automatic sentence segmentation and lexical analysis of Ancient Chinese based on BiLSTM-CRF model. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 52–58, Marseille, France. European Language Resources Association (ELRA).

C. Yan, Q. Su and J. Wang, "MoGCN: Mixture of Gated Convolutional Neural Network for Named Entity Recognition of Chinese Historical Texts," in IEEE Access, vol. 8, pp. 181629-181639, 2020, doi: 10.1109/ACCESS.2020.3026535.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fuller, M. A. (2020). The China Biographical Database User's Guide. https://projects.iq.harvard.edu/files/cbdb/files/users_guide_20200316.pdf.

Han, X., Wang, H., Zhang, S., Fu, Q., and Liu, J. S. (2019). Sentence segmentation for Classical Chinese based on LSTM with radical embedding. *The Journal of China Universities of Posts and Telecommunications*, 26(02):1–8.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Long Y., Xiong D., Lu Q., Li M., Huang CR. (2016) Named Entity Recognition for Chinese Novels in the Ming-Qing Dynasties. In: Dong M., Lin J., Tang X. (eds) Chinese Lexical Semantics. CLSW 2016. Lecture Notes in Computer Science, vol 10085. Springer, Cham. https://doi.org/10.1007/978-3-319-49508-8_34

Pang, W., Hui, C., and Chen, S. (2018). From text to data: Extracting posting data from Chinese local gazetteers. *Journal of Digital Archives and Digital Humanities*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Tsui, L. H. (2017). New approaches in digitizing Tang biographical data for the China Biographical Database. *Historical Review of Tang and Song Dynasties*, 3:20–32.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Yu, J., Wei, Y., and Zhang, Y. (2019). Automatic ancient chinese texts segmentation based on bert. *Journal sof Chinese Information Processing*, 33(11):57.

Conflict of interest statement

We declare that we have no financial or personal relationships with people or organizations that could inappropriately influence our work. There is no professional or personal interest of any kind in any product, service or company that could be construed as influencing the information presented in this manuscript.

[i] Guo Zhi was a Provincial Graduate in Taihe county, Jiangxi province. He served as a local official in Jiajing in the 2nd year. Later he was promoted to the prefect of the Songming subprefecture, and this record is also in *Ming Huan*.

Dong Jie was born in Ganyu county, Nanzhi province. He was an Imperial Academy student and served as a local official in Jiajing in the 7th year.

Wang Jiao was a Provincial Graduate in Taihe county, Jiangxi province. He served as a local official in Jiajing in the 9th year. He was known for being free of corruption and died while in office.

Deng Wenxian was a Provincial Graduate in Xinhui county, Guangdong province. He was relegated to Censor and served as a local official; he was later transfered to Tong Zhi in Huaian county.

Hu Gang was born in Wuhu county, Nanzhi province. He was an Imperial Academy student and served as a local official in Jiajing in the 13th year.

Fu Ding was a Provincial Graduate in Jinxian county, Jiangxi province. He served as a local official in Jiajing in the 14th year.

Yin Ji was an Imperial Academy student in the Fujian province and served as a local official in Jiajing in the 15th year.

Sun Ben was a Provincial Graduate in Tongzhou county and served as a local official in Jiajing in the 17th year.

Lu Jing was a Provincial Graduate in Huatin county, Nanzhi province. He had a pseudonym, Xihu. He served as a local official in Jiajing in the 22nd year.

Huang Yannian was a Provincial Graduate in Nanhai county, Guangdong province. He served as a local official in Jiajing in the 23rd year.

Jing Yicheng was born in Jiande county and was an Imperial Academy student. He served as a local official in Jiajing in the 24th year. He was skilled in poems and prose. Later he was transferred to service in Jingan county.

Zhu Jingzhi was a Provincial Graduate in Xiajiang county, Sichuan province. He served as a local official in Jiajing in the 26th year. He was well-known for being free from corruption and was later transferred to service in Maoming county.

Lin Zhi was a Provincial Graduate in Dongguan county, Guangdong province. He served as a local official in Jiajing in the 29th year and was then promoted to prefect of Dongping subprefecture.