International Journal of Geo-Information

MDPI

Article

Spatio-Temporal Information Extraction and Geoparsing for Public Chinese Resumes

Xiaolong Li [1,2], Wu Zhang [1,3,*], Yanjie Wang [4], Yongbin Tan [1,2] and Jing Xia [5]

[1] Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China; lixiaolong@ecut.edu.cn (X.L.); ybtan@ecut.edu.cn (Y.T.)
[2] CNNC Engineering Research Center of 3D Geographic Information, East China University of Technology, Nanchang 330013, China
[3] No. 325 Geological Team, Bureau of Geology and Mineral Resources of Anhui Province, Huaibei 235000, China
[4] Jiangxi Geological Museum, Nanchang 330002, China; dzbwgszhs@163.com
[5] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; xiashurang@whu.edu.cn
* Correspondence: 2020120290@ecut.edu.cn

Abstract: As an important carrier of individual information, the resume is an important data source for studying the spatio-temporal evolutionary characteristics of individual and group behaviors. This study focuses on spatio-temporal information extraction and geoparsing from resumes to provide basic technical support for spatio-temporal research based on resume text. Most current studies on resume text information extraction are oriented toward recruitment work, such as the automated information extraction, classification, and recommendation of resumes. These studies ignore the spatio-temporal information of individual and group behaviors implied in resumes. Therefore, this study takes the public resumes of teachers in key universities in China as the research data, proposes a set of spatio-temporal information extraction solutions for electronic resumes of public figures, and designs a spatial entity geoparsing method, which can effectively extract and spatially locate spatio-temporal information in the resumes. To verify the effectiveness of the proposed method, text information extraction models such as BiLSTM-CRF, BERT-CRF, and BERT-BiLSTM-CRF are selected to conduct comparative experiments, and the spatial entity geoparsing method is verified. The experimental results show that the precision of the selected models on the named entity recognition task is 96.23% and the precision of the designed spatial entity geoparsing method is 97.91%.

Keywords: named entity recognition (NER); resume information extraction; geoparsing; natural language processing (NLP); deep learning

Citation: Li, X.; Zhang, W.; Wang, Y.; Tan, Y.; Xia, J. Spatio-Temporal Information Extraction and Geoparsing for Public Chinese Resumes. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 377. https://doi.org/10.3390/ijgi12090377

Academic Editor: Wolfgang Kainz

Received: 13 June 2023
Revised: 31 August 2023
Accepted: 1 September 2023
Published: 13 September 2023

1. Introduction

With the development of the Internet and information technology, almost all the interactive information in people's daily life is transmitted on the Internet, and the amount of text information on the Internet is increasing and growing geometrically. Resumes, which carry key information about persons in textual form, are also increasing in number on the web. Statistical analyses show that well-known third-party e-job portals upload more than 300 million resumes per year [1]. A resume is a standardized, logical written expression that contains a brief description of an individual's basic information, experience, strengths, hobbies, and other relevant information.

Information extraction (IE) is the process of converting unstructured text into structured data containing information of interest [2]. Current information extraction tasks are divided into three main approaches: rule-based methods [3], machine learning methods [4], and deep learning methods [5]. The development of IE makes it possible to automated

extraction, classification and recommendation of resume [6,7], which greatly improves the efficiency of recruiters in selecting suitable job applicants. Meanwhile, resumes contain abundant temporal and spatial information, which is important for studying the spatial mobility of individuals, as well as the characteristics of the spatio-temporal distribution of a particular group of people. However, current studies on resume information extraction have not paid enough attention to temporal and spatial information and have ignored in-depth information such as personal growth path and common characteristics in resumes [8], which is a waste of the value of resume information.

Geoparsing is a cornerstone of many geographic information applications and a difficult natural language processing task [9]. It contains two important processes: geotagging and geocoding [10]. Geotagging is a special case of named entity recognition, which is aiming to identify the place names containing the geographical information from unstructured texts. Geocoding is the process of establishing the consistency between geographic coordinates and a given address. Considering the needs of the study, in this paper, we just obtain the province and city of the entity containing location information during the geocoding phase, and use the geographic coordinates corresponding to its city as the geographic coordinates of the entity.

In this paper, taking the teachers' resumes of key universities in China as an example, we explore a process and method of spatio-temporal information extraction from teachers' resumes, and design a spatial entity geoparsing method to realize the extraction and location of spatio-temporal information from resumes, which provide technical supports for spatio-temporal analysis based on resume texts and help for mining the deep information contained in resumes.

2. Related Works

2.1. Information Extraction

Resume information extraction is an important application of textual information extraction techniques, aiming to automatically extract information of interest from resume text. In studies of English resume information extraction, Ciravegna et al. [11] extracted the name, street, city, province, email, phone, fax number, and zip code information from English resumes by a rule-based approach. Kopparapu et al. [12] presented a system that can handle multiple types and formats of resumes and created an electronic database. Bodhvi et al. [13] used a semi-supervised deep learning method to parse the education section of resumes. Rakhi et al. [6] designed a resume analysis and recommendation system using NLP techniques with the objective of simplifying the employment process. In the early studies on Chinese resume information extraction techniques, rule-based methods were mainly used, such as Qiao et al. [14] researched and developed a character information extraction system based on a rule-based approach to achieve automatic extraction of semi-structured character attributes; Li et al. [15] conducted a study on encyclopedic character attribute extraction algorithms; Yu et al. [16] proposed an attribute extraction method based on remote supervision and pattern matching to extract specified person's title attributes. Since the acquisition of rules usually requires specialized domain knowledge, the generalization ability of rule-based methods is low. Subsequently, machine learning methods for entity extraction have emerged. Dong et al. [17] proposed a method for extracting key information from teachers' homepages based on conditional random fields (CRFs); Chen et al. [18] proposed a "two-step" resume information extraction algorithm by combining the syntactic information of resumes with the design of "Writing Style", and achieved the accurate extraction of resume information without defining rules and annotating data. In recent years, with the development of artificial neural networks, deep learning methods have also been widely used in resume information extraction [19]. Some scholars have built named entity recognition models for Chinese e-resumes based on the BERT language model, which shows good performance [8,20].

The task of spatio-temporal information extraction focuses on identifying and extracting temporal and spatial information from text data, and constructing relationships

between temporal and spatial information in order to describe the changes in the spatial location of the research object within a certain time period, so as to explore the intrinsic patterns and characteristics of the research object's behaviors. Some scholars have explored the extraction of temporal information in Chinese texts from the perspective of linguistics, mainly by analyzing the constituent elements of time and the form of time word composition in Chinese and adopting the concept of temporal expressions to achieve the identification [21,22]. Based on the extraction of time elements, the normalized expression of time phrases is achieved by defining the types of time relations in Chinese descriptions and parsing the internal rules of time expressions [23]. By summarizing the characteristics of time information description in Chinese texts, Zhang et al. [24] constructed a time lexicon and a time description pattern library, and designed algorithms for the normalized expression of time information and semantic inference. Qiu et al. [25] conducted the extraction of temporal information in geological reports by constructing a temporal gazetteer. For the extraction of spatial entities, the extraction of place name (or toponym) is the main focus. The existing methods for recognizing toponyms can be divided into three types: rule-based methods, machine learning methods, and deep learning methods. The rule-based methods are to build gazetteers, to combine the characteristics of the word composition and lexical features of a toponym, and to generalize the general rules of place name expressions for the recognition of toponym entities. This approach has the advantages of simplicity and precision, but the limitations of the constructed gazetteer cannot handle situations such as new place names and complex syntax. The machine learning approach does not require specialized language knowledge, is more robust and flexible than the rule-based approach. In recent years, machine learning models such as the hidden Markov model (HMM), the Support Vector Machine, the maximum entropy Markov model (MEMM), and CRFs [26,27] have been used for the recognition of toponyms. With the rise and development of deep learning, toponym entity recognition methods based on deep learning models have also been widely used, and the classical models are BiLSTM-CRF [28], BERT-CRF [29], etc.

2.2. Geoparsing

Gritta et al. [10] conduct a detailed geoparsing survey, which evaluates and analyzes the performance of a number of leading geoparsers on a number of corpora and highlight the challenges in detail. For the aim of obtaining the geographic content of the social media message, Gelernter et al. [30] present a method to geo-parse the short, informal messages known as microtext.

For Chinese geotagging, the method of neural network is mostly used, and the mainstream method is based on a pre-training model [31,32]. Chinese geocoding mostly relies on map service platforms, and the mainstream Chinese map service platforms are: Baidu Map (https://lbsyun.baidu.com, accessed on 12 June 2023), Gaode Map (https://lbs.amap.com, accessed on 12 June 2023) and Tianmap (http://lbs.tianditu.gov.cn, accessed on 12 June 2023), etc. Due to the different geocoding rules and data sources provided by different map service providers, the results of geocoding will be different. He et al. [33] fused and optimized multi-source online coding services to reduce the result bias caused by geocoding differences and improve the efficiency of geocoding work. Zhu [34] used four mapping platforms to conduct the geocoding of some address data, comparing the geocoding errors for community addresses and road addresses. In this paper, for toponyms geoparsing, we not only used the map service platforms, but also used web encyclopedic knowledge, expanding the information sources for geoparsing.

3. Materials and Methods

3.1. Data

The public resumes of university teachers are usually posted on the official websites of universities, which are generally easy to find. In this study, we obtained the teacher resumes from 35 "Project 985" universities in China. The 35 "Project 985" universities are

listed in Appendix A. Then, we collected 51,438 resumes from the official websites of these universities. After cleaning, 28,306 resumes were remained.

The university teacher resumes are semi-structured texts, and the writing style has certain regularity. Figure 1 shows an example of crawled resumes of key university teachers, from which it can be seen that the content of the teacher resume can be divided into modules, such as educational experience, work experience, academic positions, and teaching courses. Browsing through a large number of teacher resumes, we can see that they also contain modules on teaching and research, awards and honors, papers and achievements, etc. These modules are usually identified by keywords such as "basic information", "educational experience", "work experience", and "research interests", i.e., each module appears in the form of "caption keyword + module content" [35]. Therefore, it is particularly important to make good use of these caption keywords in the teachers' resumes in order to effectively chunk the content of the resumes.


```
{
    "name": "Li Yulong",
    "resume": [
        "Li Yulong",
        "Email：\tyulongli@pku.edu.cn",
        "Professional Title：profeesor",
        "Office Address：No.5,Summer Palace Road, Haidian District, Beijing, China,
        Peking University, Lv Zhihe Building, 100871",
        "Lab Address：No.5, Summer Palace Road, Haidian District, Beijing, China,
        Peking University, Lv Zhihe Building, 100871",
        "Laboratory homepage:http://www.yulonglilab.org",
        "Personal Profile",
        "Research Areas",
        "Representative Papers",
        "Educational Experience:",
        "2000-2006, PhD, Department of Neurobiology, Duke University",
        "1996-2000, B.S., College of Life Sciences, Peking University",
        "Honour Awards:",
        "Scientific Exploration Award for Life Sciences，2019",
        "National Foundation for Distinguished Youth Science，2019",
        "The 12th Tan Jiazhen Prize for Innovation in Life Sciences，2019",
        "Zhang Xiangtong Young Scientist Award in Neuroscience，2019",
        "China's Top 10 Medical Technology News，2018",
        "Ten Major Advances in Life Sciences in China，China Association for Science and
        Technology Consortium of Life Science Societies，2018",
        "Innovative Breakthrough Award of the Zhongyuan Xiehe Life Medicine Prize，2018",
        "Boehringer Ingelheim Researcher Award，2018",
        "Greenleaf Biomedical Outstanding Young Scholar Award，2015",
        "Work Experience:",
        "2020-present, Professor, School of Life Sciences, Peking University",
        "2012-present, Research Fellow, School of Life Sciences, Joint Centre for Life Sciences,
        McGovern Institute for Brain Research, Peking University",
        "2019-2020, Associate Professor, School of Life Sciences, Peking University",
        "2012-2019, Assistant Professor, School of Life Sciences, Peking University",
        "2006-2012, Postdoctoral Fellow, Department of Molecular and Cellular Physiology,
        Stanford University",
        "Academic Positions:",
        "2019-present，JournalofNeurochemistry，Editorial Board",
        "2018-present，Chinese American Society for Biological Sciences，Member",
        "2001-present，American Society for Neuroscience，Member",
        "Teaching Courses:",
        "Genetics Seminar",
        "Integrated Science Experiment Classes",
        "Life Sciences Intensive Challenge Class Literature Discussion Session",
        ......
    ]
}
```


Figure 1. Example of crawled resumes of key university teachers. The resume has been shortened for display.

3.2. Methodology

Figure 2 illustrates the framework of the proposed method which is divided into four parts: data acquisition, caption lexicons construction, information extraction and geocoding. In the data acquisition section, 28,306 valid teacher resumes were collected from the official websites of 35 key universities in China using web crawler technology and stored in JSON data format. In the caption lexicons construction section, a caption lexicon of teacher resumes from each key university was obtained based on statistical and text similarity calculation methods. In the information extraction section, the target entities

in the university teacher resumes were identified by the constructed information extraction scheme. In the geocoding section, a spatial entity geocoding method is designed to obtain the geographical coordinates of toponyms.

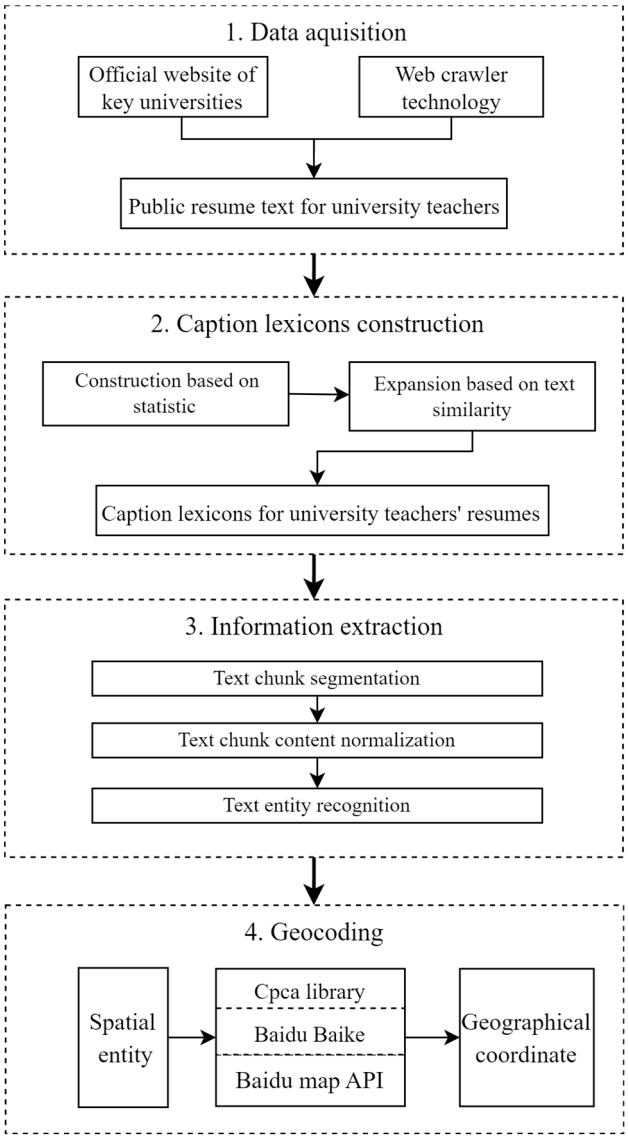

Figure 2. Framework of spatio-temporal information extraction and geocoding based on public resumes.

4. Resume Caption Lexicons Construction

As mentioned in Section 3.1, the modules in a teacher resume usually take the form of "caption keyword + module content". Using these headwords to chunk the content of a resume is a prerequisite for subsequent information extraction. In this study, a statistical-based approach is used to initially obtain a high-frequency caption lexicon, and then a text similarity-based approach is used to expand the caption lexicon to obtain a comprehensive caption lexicon for each university, as each university's resume style is not consistent.

4.1. Statistical-Based Caption Lexicon Construction

Browsing through a large number of teacher resumes, we can see that the caption keywords are generally separated and are generally between four and seven in length, while they may end in a colon. Through these characteristics, rules were established to initially count the eligible caption words and their word frequencies, and finally the caption words ranked in the top 15 in terms of word frequency for each university were retained.

By summarizing and classifying the results of the preliminary statistics of caption words from all universities, we divided the caption words of teacher resumes into 11 categories, as shown in Table 1.

Table 1. The construction of university teacher resumes caption lexicon based on statistics.

Caption Category	Caption Lexicon
Basic Information	"Basic Information", "Personal Introduction", "Personal Information"
Personal Resume	"Personal Resume", "Curriculum Vitae", "Study and Work Experience"
Work Experience	"Work Experience", "Professional Experience", "Job Resume"
Education Experience	"Educational Experience", "Educational Background", "Study Experience", "Academic Experience"
Research Areas	"Research Areas", "Research Content", "Research Direction"
Teaching and Research	"Lecture Courses", "Teaching Course", "Teaching Situation", "Teaching", "Teaching and Scientific Research", "Teaching Curriculum"
Awards and Honors	"Scholastic Honor", "Awards and Honors"
Tenure and Part-Time	"Adjunct Research Position", "Social Position", "Academic Participation"
Scientific Research	"Scientific Research", "Research Projects", "Hosting Projects", "Research Summary"
Thesis Results	"Published Papers", "Books and Papers", "Representative Researches", "Research Results", "Representative Papers", "Academic Works"
Contact Details	"Contact Details"

4.2. Text Similarity-Based Caption Lexicon Expansion

Text similarity is the degree of similarity between texts and measures the commonality and difference between texts. There are many ways to calculate text similarity, and this paper focuses on two methods: word vector and edit distance. The essence of word vector is to embed words into a low-dimensional vector representation, which makes it possible for semantically similar words to be nearer in spatial distance, and also facilitates the calculation of similarities between words. In addition, this representation of words is a good solution to the problem of semantic deficit and dimensional disaster caused by the bag-of-words model due to the independence of words [36]. The edit distance, also known as the Levenshtein Distance, is a string-level measure of text similarity, measured by the number of operations required to transform a string into another string, including insertion, modification and deletion.

In this paper, we use a combination of word vector and edit distance to measure text similarity to discover new caption words. The formula for calculating the text similarity between short texts is as follows:

$$Sim_{a,b} = \alpha * \cos(S_a, S_b) + \beta * (\frac{1 - ED_{a,b}}{\max(L_a, L_b)}) \quad (1)$$

where S_a, S_b is the weighted average of the word vectors of the short text a, b after splitting and deactivating stopwords, respectively. L_a, L_b is the length of the short text a, b after deactivating the stopwords, respectively. $ED_{a,b}$ is the edit distance of the short text a, b. α, β are the weighting factors used to adjust the two factors. Through comparative experiments, we got the optimal weighting factors that α is 0.8 and β is 0.2.

Based on the constructed text similarity measure formula, the teacher resume caption lexicon constructed by the statistical-based method was expanded to obtain teacher resume caption lexicons for each university.

5. Resume Information Extraction Scheme

The resume information extraction scheme for university teachers constructed in this study is divided into three main steps, namely, resume text chunk segmentation, resume text chunked content normalization and resume text entity recognition. Among them, the purpose of resume chunk segmentation is to divide a resume into different subject chunks by resume caption words, and each chunk describes the same topic, such as "educational experience" and "work experience". The chunked resume text has a messy format and cannot be used as input for the named entity recognition model. The purpose of the normalization process is to standardize the content of the chunked resumes into the input form for the named entity recognition model. For the task of recognizing entities, we focus on the recognition of temporal and spatial entities in resumes. For the temporal entity, we use a rule-based approach for pattern recognition; for the spatial and other entity, a deep learning approach is used.

5.1. Resume Text Chunking

We adopt a rule-based pinch-force cutting method to achieve the segmentation of resume text chunks, and the steps are as follows:

Step 1: Caption trigger word targeting

According to the constructed caption lexicon of teacher resumes of each university, rules are established to match the caption words in the resume text. The matched caption words are assigned to the corresponding categories, while the position indexes of the caption words are noted for chunking.

Step 2: Caption trigger word sorting

Sorting the caption trigger words by the position indexes from smallest to largest, getting an ordered sequence of caption trigger words.

Step 3: Resume text cutting

Firstly, according to the ordered sequence of caption trigger words, based on the rule that the content between two trigger words belongs to the previous trigger word, the start and end position indexes of the resume text under the resume caption category to which the caption trigger word belongs are obtained. Then, the resume texts are chunked according to the position indexes. Lastly, the text chunks are classified. As we focus on the spatial and temporal information, only four categories of resume chunks, namely "basic information", "personal resume", "educational experience" and "work experience", are remained.

The above resume text chunking algorithm relies heavily on the caption lexicons, implemented with the rule that the content between two trigger words belongs to the previous trigger word. In this paper, a more complete caption lexicon of teacher resumes for each university is obtained, and therefore this resume text chunking algorithm performs better in this study.

5.2. Resume Text Chunk Content Normalization

By resume text chunking, teacher resumes are divided into different topics according to caption words, which prepares for the subsequent information extraction. However, some of the chunked resumes are not suitable to be the input of the named entity extraction model, and need to be normalized into line-by-line resume descriptions. For example, the education section in a chunked resume is like this: "'2014–2018', 'Doctor', 'Astrophysics', 'Southern Methodist University', '2011–2013', 'Master', 'Photogrammetry and Remote Sensing', 'Wuhan University', '2007–2011', 'Bachelor', 'Geographic Information System', 'China University of Geosciences, Wuhan'". It has to be normalized as: "'2014–2018 Doctor

Astrophysics Southern Methodist University', '2011–2013 Master Photogrammetry and Remote Sensing Wuhan University', '2007–2011 Bachelor Geographic Information System China University of Geosciences, Wuhan'".

In order to automate the normalization of chunked resume text to obtain line-by-line resume descriptions, we define rules for the merging of resume text data as Table 2.

Table 2. The rules and operations for university teacher resumes normalization.

Rules	Operations
Time	Merge with next line
End with punctuation	Merge with next line
Begin with punctuation	Merge with previous line
Short text	Merge with previous line
Plain English text	Merge with previous line

5.3. Resume Text Entity Recognition

5.3.1. Rule-Based Temporal Entity Recognition

The time information in resumes has a distinct writing pattern, and a rule-based approach is used to write regular expressions for pattern matching recognition. Browsing through the time descriptions in the teacher resumes of various universities, their writing types were summarized as shown in Table 3. Considering the needs of this study, we only extract year information. Additionally, for the experience without year information, omission was done.

Table 3. Summary of time writing types in university teachers' resumes.

Type of Time Description	Example
Full description	06/2012–11/2017 2011.2–2011.8 2017/07–2020/10 07/2002–08/2005 1993–1996
Omitted description	98/09–01/06 09/94–07/97 04–13

The regular expressions for temporal entity extraction, which are written by Python using the Re (https://docs.python.org/3/library/re.html, accessed on 12 June 2023), are as follows:

- P1: re.compile(r'\d{4}'),
- P2: re.compile(r'(\d{2})/'),
- P3: re.compile(r'/(\d{2})'), and
- P4: re.compile(r'\d{2}').

The P1 directly identifies the year in the full time description, e.g., "2012" and "2017" in "June 2012–November 2017" P2, P3 and P4 are used to extract the year in the omitted time description. Additionally, for the omitted time description like "98/09–01/06", a judgement and normalization process is required to ensure that the correct and standardized year entities are extracted.

5.3.2. Deep Learning-Based Entity Recognition

Non-temporal entities such as places, institutions and positions are represented in various forms and with little regularity, thus we use deep learning methods to extract such entities. In this paper, BERT-BiLSTM-CRF is used as the resume entity recognition model.

BERT uses a bi-directional Transformer [37] encoder and a Masked Language Model to implement a bi-directional language model, which is pre-trained to obtain a representation

of each word considering contextual information. Compared to traditional language models, BERT has stronger representational power and is able to solve problems such as multiple meanings of words. The Self-attention Mechanism is the main module of the Transformer encoder used by BERT, which uses the Query-Key-Value (QKV) model to map each input word into three different spaces to obtain the query vector Q, the key vector K and the value vector V, respectively. The scoring value is obtained by calculating Q and K with the dimension of the input vector d_K, and then the scoring value is calculated with V to finally obtain a new representation of each word considering the inter-word relationships, as shown in Equation (2).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \tag{2}$$

LSTM (Long Short-Term Memory) is a variant of recurrent neural network (RNN), which can effectively solve the gradient explosion or disappearance problem of simple recurrent neural networks [38]. The LSTM network introduces a new internal state dedicated to linear recurrent information transfer, while the input gate i_t, the forgetting gate f_t and the output gate o_t are introduced through a gating mechanism to control the path of information transfer in order to control the updating, transferring and forgetting of information in the memory unit. The computational formula of the LSTM network can be succinctly described as follows:

$$\begin{bmatrix} \widetilde{c}_t \\ o_t \\ i_t \\ f_t \end{bmatrix} = \begin{bmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{bmatrix} \left(W \begin{bmatrix} x_i \\ h_{t-1} \end{bmatrix} + b \right) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \tag{4}$$

$$h_t = o_t \odot \tanh(c_t) \tag{5}$$

where σ is the Sigmoid activation function; o_t, i_t and f_t are the output, input and forgetting gates, respectively; W and b are the net parameters; $\odot$ is the element-by-element multiplication; and h_t represents the output of the hidden layer memory unit at the moment of t.

BiLSTM (Bi-directional Long Short-Term Memory) obtains preceding and following text information through LSTM in both directions, thus compensating for the deficiency that unidirectional LSTM cannot learn the following information. However, the individual states of the output sequence of the BiLSTM network are independent of each other, so the problem of incoherent entity labels in the sequence annotation will arise, which requires the introduction of label-to-label constraint relations in the prediction process.

CRFs are conditional probability distribution models for outputting random variables given a set of input random variables, which takes into account the interdependencies between the labels and ensures that the output labels are reasonable. Therefore, the inclusion of a CRF layer after the output of a BiLSTM network allows the prediction results to be constrained. In the application of labelling problems, linear-chain conditional random fields are often used. Given an input sequence X, and assuming that training yields the corresponding output label sequence Y, the formula is as follows:

$$P(Y|X) = \frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K} \omega_k f_k(y, x) \right) \tag{6}$$

$$Z(x) = \sum_{y} \exp\left(\sum_{k=1}^{K} \omega_k f_k(y, x) \right) \tag{7}$$

where ω denotes the weight vector, f denotes the feature function, and $Z(x)$ is the normalization factor.

6. Spatial Entities Decoding

The spatial entity, i.e., an entity containing spatial location information, mainly include two types of entities, namely places and organizations. The decoding of spatial entities in this study is to determine the information of the administrative division of the province and city where the given spatial entity is located, and then obtain the geographical coordinates corresponding to the administrative division to be used as the spatial coordinates of the entity. For foreign spatial entities, we locate them to the national scale, e.g., Oxford University is located to the UK. Then, the OSM (Open Street Map) latitude and longitude coordinates corresponding to the country are obtained through the Nominatim API (https://nominatim.org/release-docs/latest, accessed on 12 June 2023) to achieve the geocoding of foreign spatial entities. We use a combination of Baidu Baike (https://Baike.baidu.com, accessed on 12 June 2023) search, Baidu Map Application Programming Interface (API) (https://lbsyun.baidu.com/faq/api?title=webapi/guide/webservice-placeapi, accessed on 12 June 2023), and cpca (https://github.com/DQinYuan/chinese_province_city_area_mapper, accessed on 12 June 2023) library mapping to achieve the geocoding of spatial entities.

For the included entry, the Baidu Baike returns a normative entry page. The entry page can be broadly divided into four parts: title, overview, basic information column and body content, as shown in Figure 3. By parsing these contents, the location of the spatial entity can be obtained. For entries that have not yet been included, Baidu Baike will return content-related entries for users' reference. The Baidu Map Location Retrieval Service API can be queried to obtain the province and city where the input keyword is located. The cpca library in Python can extract the province, city and district from strings and perform mapping, which can quickly identify the province and city expressed in spatial entity strings.

Figure 3. Example of Baidu Baike page content division.

6.1. Spatial Entity Geocoding Based on Baidu Baike

The method based on the Baidu Baike is the key research content. In this method, for the entry included in Baidu Baike, it obtains the location of spatial entities by web page parsing; for the entry that have not yet been included in Baidu Baike, it will determine whether the returned related entry and the retrieved entity are the same entity, and if they are, parsing the page to obtain location. The algorithm flow of this method is shown in Figure 4. The "web page parsing" in the algorithm flow is the parsing of the title, overview, basic information column and body content. Algorithm 1 (page parsing algorithm) for web page parsing is as follows:

Algorithm 1. (page parsing algorithm)

Step1: Recognize the current webpage title through the cpca library, and judge the recognition result. If the province and city is complete, Step5 will be executed, otherwise Step2 will be executed.
Step2: Iterate through the basic information column of the webpage and identify the content of the location. Execute Step5 if the province and city (city may be missing) is obtained, otherwise execute Step3.
Step3: Pattern matching to identify the location in the web paragraph. Execute Step5 if province and city (city may be missing) is obtained, otherwise execute Step4.
Step4: Do word segmentation for the webpage overview and body content, and build a dictionary of words in the form of {key = city name, value = number of occurrences}. If the dictionary is not empty and the maximum number of occurrences is not less than 5, the city with the highest number of occurrences will be the final result and the corresponding province will be obtained; otherwise, the province and city will be empty. Execute Step5.
Step5: Finish and return to the province and city.

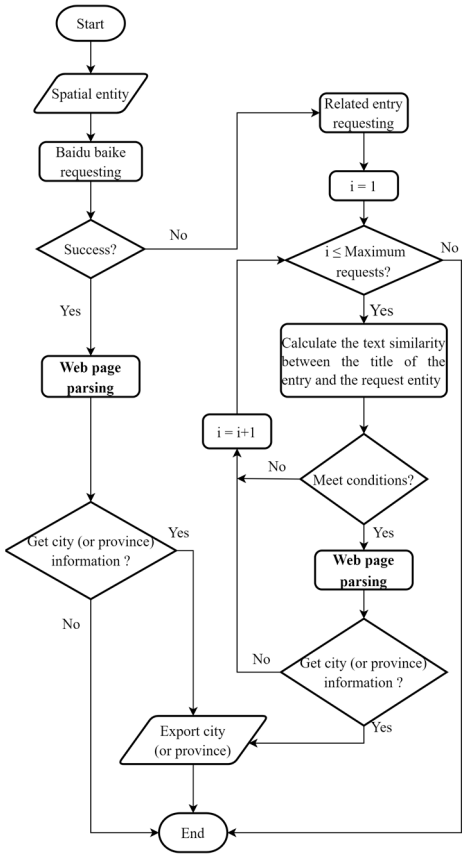

Figure 4. Method of spatial entity geocoding based on Baidu Baike.

6.2. Threshold Setting

In the spatial entity geocoding method based on Baidu Baike and Baidu Map API, the returned relevant entry titles and POI names may not be the same entity as the search keywords, so threshold settings are required to improve the accuracy of spatial entity geocoding. Based on the actual extraction situation, the final threshold settings are described in Table 4.

Table 4. Threshold settings for the spatial entity geocoding method.

Method	Description of the Threshold Settings
Baidu Baike Search	When the text editing distance between the retrieved entity and the title of the related entry after deactivating stopwords is within 5 or the cosine similarity at the character level [39] is greater than or equal to 0.9, the retrieved entity is considered to be the same entity as the related entry if it appears in the overview paragraph of the entry or if the cosine similarity is greater than or equal to 0.95.
Baidu Map API Query	After deactivating stopwords for the POI names returned by the Baidu Map Location Retrieval Service API, if the text editing distance with the retrieved entity is within 2 or the cosine similarity is greater than or equal to 0.95, the returned POI is considered to be the same entity as the retrieved entity.

7. Experimental Evaluation

7.1. Dataset

In this paper, we use a Chinese resume dataset (Resume NER), presented by Zhang et al. [40] in 2018, as the database to train the resume entity recognition model BERT-BiLSTM-CRF. This dataset is crawled from the Sina Finance website and consists of 1027 randomly selected resume summaries of executives of listed companies in the Chinese stock market, annotated with 8 types of named entities using the YEDDA system, namely nationality, education, native place/location, name, organization, profession, ethnicity, and position. The total number of sentences in the Resume NER dataset is over 4700, and the training set, validation set and test set are divided according to 8:1:1.

7.2. Evaluation Indexes

In this study, the precision P, recall R and F1 values are selected as model and geocoding method evaluation indexes.

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$\mathrm{F1} = \frac{2PR}{P + R} \tag{10}$$

where TP, FP, and FN is the number of true positive samples, false positive samples, and false negative prediction samples, respectively.

7.3. Recognition Results

The recognition results of the BERT-BiLSTM-CRF model on the test set are shown in Table 5. To demonstrate the superiority of the model, three baseline models were selected, and the evaluation metrics of each model on the test set are shown in Table 6. It can be seen that the BERT-CRF and BERT-BiLSTM-CRF models are significantly more effective than the BiLSTM and BiLSTM-CRF models, indicating that pre-training on a large-scale text corpus can effectively improve the model performance on downstream tasks. The sequence modelling capability of CRF constrains the model output, which can also improve the prediction of models to a large extent. The BERT- BiLSTM-CRF model combines BERT

with BiLSTM to perform multi-layer feature extraction on the input, and then constrains the model output by a CRF, which ultimately achieves the best results for all type entities and outperforms the other comparison models in terms of precision and F1 value.

Table 5. Evaluation of various entity recognition results of the BERT-BilSTM-CRF model.

Type of Entity	Precision/%	Recall/%	F1 Value/%
Nationality	100	100	100
Education	99.11	99.11	99.11
Native place/location	100	100	100
Name	100	100	100
Orgnization	94.07	94.65	94.36
Profession	97.06	100	98.51
Ethnicity	100	100	100
Position	96.39	94.47	95.42

Table 6. Evaluation indexes of each model on the test set.

Models	Precision/%	Recall/%	F1 Value/%
BiLSTM	90.69	87.30	88.97
BiLSTM-CRF	91.47	88.77	90.10
BERT-CRF	90.10	95.70	95.63
BERT-BiLSTM-CRF	96.23	95.63	95.93

7.4. Geocoding Results

From the resume entities, 631 spatial entities are randomly selected, and the geographical locations of the entities were manually labeled as test data. Then, the test entities are geocoded by the spatial entity geocoding method. The results show that in the 621 entities extracted with location, 608 were correctly located, and the specific index results are shown in Table 7.

Table 7. Result of test experiment.

Methods	Precision/%	Recall/%	F1 Value/%
Spatial entity geocoding	97.91	96.35	97.12

The method of geocoding is generally good and can meet the needs of the study. However, for some abolished or renamed spatial entities such as the "Twenty-ninth Research Institute of the State Ministry of Machinery and Electronics Industry", "Fifth Institute of China Aerospace Industry Corporation", "Factory 230 of the Aerospace Corporation", the locations are difficult to obtain and easily misrecognition.

8. Conclusions

Resumes contain rich spatio-temporal information about people's behaviors, which is of great value for studying the spatio-temporal evolutionary characteristics of individual and group behaviors. However, current research on resume information extraction lacks attention to spatio-temporal information and fails to fully explore the analytical value of resumes. In order to make full use of the spatio-temporal information of resumes and provide technical support for the spatio-temporal analysis research based on resumes, this study proposes a spatio-temporal information extraction and geoparsing method for people's resumes by combining NLP and geoparsing techniques, which effectively realizes the extraction and geoparsing of spatio-temporal information in resumes. The method consists of three major aspects: combining statistical methods and text similarity calculation methods to construct the title thesaurus of teachers' resumes in various colleges and universities; realizing the recognition of target entities in teachers' resumes through

the designed resume information extraction solutions; and implementing the positioning of spatial entities in teachers' resumes through the constructed spatial entity geocoding method. Experiments show that the named entity recognition model selected in this paper is significantly better than other models, and the constructed spatial entity geocoding method has higher accuracy, which can provide support for the research of spatio-temporal analysis based on resume data.

At the same time, there are some shortcomings in this study. On the one hand, the spatial entity location method mainly relies on Baidu Baike knowledge, which is a single source of information and is task specific to a certain extent, and the generalization ability needs to be further verified. Future research can explore more relevant entity information sources to improve the accuracy and generalization of the spatial entity location method. On the other hand, in the work of extracting resume information of college teachers, the named entity recognition method used in this paper takes the canonical resume information item by item as the input, so a certain amount of manual effort needs to be invested in the text chunk normalization. Subsequent research can consider adopting the event extraction method to achieve a fully automatic extraction of teacher resume information to reduce manual consumption in text content normalization.

Author Contributions: Conceptualization, Xiaolong Li and Wu Zhang; data curation, Wu Zhang and Jing Xia; formal analysis, Xiaolong Li and Wu Zhang; methodology, Xiaolong Li, Wu Zhang and Jing Xia; supervision, Xiaolong Li and Yongbin Tan; writing—original draft, Xiaolong Li and Wu Zhang; writing—review and editing, Yanjie Wang and Yongbin Tan. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 42261078), the Jiangxi Provincial Key R&D Program (grant number 20223BBE51030) and the Science and Technology Research Project of Jiangxi Bureau of Geology (grant number 2022JXDZKJKY08).

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/jiesutd/LatticeLSTM, accessed on 11 January 2023.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

A list of the 35 "Project 985" universities: Tsinghua University, Peking University, Zhejiang University, Shanghai Jiao Tong University, Nanjing University, Fudan University, University of Science and Technology of China, Huazhong University of Science and Technology, Wuhan University, Xi'an Jiaotong University, Harbin Institute of Technology, Beijing Normal University, Beihang University, Tongji University, Southeast University, Renmin University of China, Beijing Institute of Technology, Nankai University, Shandong University, Tianjin University, Central South University, Jilin University, Northwestern Polytechnical University, Xiamen University, South China University of Technology, Dalian University of Technology, East China Normal University, China Agricultural University, Hunan University, Chongqing University, Northeastern University, Lanzhou University, Ocean University of China, Northwest A&F University, and Minzu University of China.

References

1. Zu, S.; Wang, X. Resume Information Extraction with A Novel Text Block Segmentation Algorithm. *Int. J. Nat. Lang. Comput.* **2019**, *8*, 29–48. [CrossRef]
2. Grishman, R. Twenty-five years of information extraction. *Nat. Lang. Eng.* **2019**, *25*, 677–692. [CrossRef]
3. Soderland, S. Learning information extraction rules for semi-structured and free text. *Mach. Learn.* **1999**, *34*, 233–272. [CrossRef]
4. Freitag, D.; McCallum, A. Information extraction with HMMs and shrinkage. In Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction, Orlando, FL, USA, 18–19 July 1999; pp. 31–36.
5. Yang, Y.; Wu, Z.; Yang, Y.; Lian, S.; Guo, F.; Wang, Z. A Survey of Information Extraction Based on Deep Learning. *Appl. Sci.* **2022**, *12*, 9691. [CrossRef]
6. Bharadwaj, R.; Mahajan, D.; Bharsakle, M.; Meshram, K.; Pujari, H. Resume analysis using NLP. In *Inventive Systems and Control*; Suma, V., Lorenz, P., Baig, Z., Eds.; Springer Nature: Singapore, 2023; pp. 551–561.

7. Li, X.; Shu, H.; Guang, Y.; Zhai, Y.; Yang, Z. Survey of the Application of Natural Language Processing for Resume Analysis. *Comput. Sci.* **2022**, *49*, 66–73. [CrossRef]
8. Shen, K.; Huang, H.; Hua, B. Constructing Knowledge Graph with Public Resumes. *Data Anal. Knowl. Discov.* **2021**, *5*, 81–90.
9. Tao, L.; Xie, Z.; Xu, D.; Ma, K.; Qiu, Q.; Pan, S.; Huang, B. Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 598. [CrossRef]
10. Gritta, M.; Pilehvar, M.T.; Limsopatham, N.; Collier, N. What's missing in geographical parsing? *Lang. Resour. Eval.* **2018**, *52*, 603–623. [CrossRef]
11. Ciravegna, F.; Lavelli, A. LearningPinocchio: Adaptive information extraction for real world applications. *Nat. Lang. Eng.* **2004**, *10*, 145–165. [CrossRef]
12. Kopparapu, S.K. Automatic extraction of usable information from unstructured resumes to aid search. In Proceedings of the 2010 IEEE International Conference on Progress in Informatics and Computing, Shanghai, China, 10–12 December 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 99–103.
13. Gaur, B.; Saluja, G.S.; Sivakumar, H.B.; Singh, S. Semi-supervised deep learning based named entity recognition model to parse education section of resumes. *Neural Comput. Appl.* **2021**, *33*, 5705–5718. [CrossRef]
14. Qiao, L.; Li, C.; Zhong, Z.; Wang, J.; Liu, D. Research on People's Information Extraction Based on Rules. *J. Nanjing Norm. Univ. (Nat. Sci. Ed.)* **2012**, *35*, 134–139.
15. Li, H.; Yang, Y.; Yin, H. Research on character attributes extraction based on rules from Baidu encyclopedia. *J. Integr. Technol.* **2013**, *2*, 1–4.
16. Yu, D.; Liu, C.; Tian, Y. Personal title and career attributes extraction based on distant supervision and pattern matching. *J. Comput. Appl.* **2016**, *36*, 455–459.
17. Dong, F.; Wang, J. Personal Information Extraction of the Teaching Staff Based on CRFs. In Proceedings of the International Conference on Network & Information Systems for Computers, Wuhan, China, 23–25 January 2015; IEEE: Piscataway, NJ, USA, 2015.
18. Chen, J.; Zhang, C.; Niu, Z. A two-step resume information extraction algorithm. *Math. Probl. Eng.* **2018**, *2018*, 5761287. [CrossRef]
19. Yang, Y.; Bai, Y.; Cai, D.; He, J. Information extraction for resumes of scientific and technological figures. *Comput. Eng. Des.* **2021**, *42*, 3099–3106.
20. Guo, J.; Wan, G.; Hu, X.; Wei, Z. Chinese resume named entity recognition based on BERT. *J. Comput. Appl.* **2021**, *41*, 15–19.
21. Lin, J.; Cao, D.; Yuan, C. Automatic TIMEX2 tagging of Chinese temporal information. *J. Tsinghua Univ. (Sci. Technol.)* **2008**, *48*, 117–120.
22. Wu, T.; Zhou, Y.; Huang, X.; Wu, L. Chinese time expression recognition base on automatically generated basic-time-unit rules. *J. Chin. Inf. Process.* **2010**, *24*, 3–10.
23. Wen, Y.; Tan, H.; Zheng, J. Research on time standardization based on rules. In Proceedings of the 2009 International Information Technology and Applications Forum, Chengdu, China, 15–17 May 2009; pp. 49–51.
24. Zhang, C.J.; Zhang, X.Y.; Li, M.; Wang, S. Interpretation of temporal information in Chinese text. *Geogr. Geo-Inf. Sci.* **2014**, *30*, 1–7.
25. Qiu, Q.; Xie, Z.; Wu, L.; Tao, L. Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques. *Earth Sci. Inform.* **2020**, *13*, 1393–1410. [CrossRef]
26. Wu, L.; Liu, L.; Li, H.; Gao, Y. A Chinese Toponym Recognition Method Based on Conditional Random Feild. *Geomat. Inf. Sci. Wuhan Univ.* **2017**, *42*, 150–156.
27. Mao, P.; Teng, W. Complex Chinese place name recognition based on conditional rangdom field and rule improvement. *Eng. J. Wuhan Univ.* **2020**, *53*, 447–454.
28. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
29. Xu, L.; Dong, Q.; Liao, Y.; Yu, C.; Tian, Y.; Liu, W.; Li, L.; Liu, C.; Zhang, X. CLUENER2020: Fine-grained named entity recognition dataset and benchmark for Chinese. *arXiv* **2020**, arXiv:2001.04351.
30. Gelernter, J.; Balaji, S. An algorithm for local geoparsing of microtext. *GeoInformatica* **2013**, *17*, 635–667. [CrossRef]
31. Liu, X.; Li, Y.; Yin, B.; Tian, Q. Chinese address understanding by integrating neural network and spatial relationship. *Sci. Surv. Mapp.* **2021**, *46*, 165–171+212.
32. Zhang, H.; Du, Q.; Chen, Z.; Zhang, C. A Chinese Address Parsing Method Using RoBERTa-BiLSTM-CRF. *Geomat. Inf. Sci. Wuhan Univ.* **2022**, *47*, 665–672.
33. He, C.; Wan, Y. Optimization and Application of Online Multi-source Geocoding Fusin. *Geospat. Inf.* **2023**, *21*, 45–47+116.
34. Zhu, X. Comparison of geocoding errors for community addresses and road addresses. *Jiangsu Sci. Technol. Inf.* **2022**, *39*, 70–75.
35. Yan, W. Information Extraction for Semi-Structured Chinese Resume. Master's Thesis, South China University of Technology, Guangzhou, China, 2018.
36. Chen, E.; Jiang, E. Review of Studies on Text Similary Measures. *Data Anal. Knowl. Discov.* **2017**, *1*, 1–11.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
38. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **2002**, *5*, 157–166. [CrossRef] [PubMed]

39. Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; Du, X. Analogical Reasoning on Chinese Morphological and Semantic Relations. *arXiv* **2018**, arXiv:1805.06504.
40. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. *arXiv* **2018**, arXiv:1805.02023.