Sensors and Materials, Vol. 32, No. 12 (2020) 4489–4504
MYU Tokyo

S & M 2424

Rapid Extraction of Research Areas from Scientific and Technological Literature

Chuan Yin,[1] Wanzeng Liu,[2*] Duoduo Yin,[3] Xi Zhai,[2]
Kexin Liu,[1] Changfeng Jing,[1] and He Huang[1]

[1]Beijing University of Civil Engineering and Architecture,
School of Geomatics and Urban Spatial Informatics, Beijing 102616, China
[2]National Geomatics Centre of China, Information Service Department, Beijing 100830, China
[3]Capital Normal University, College of Resource Environment and Tourism, Beijing 100048, China

(Received September 29, 2020; accepted December 8, 2020)

Keywords: smart city, knowledge extraction, study area extraction, BiLSTM-CRF, random forest model

Along with the rapid development of Internet Plus, big data, and other technologies, the construction of smart cities is promoting the transformation and upgrading of mapping geographic information models from traditional information services to intelligent services with spatial sensing. At present, however, most of the knowledge needed to provide intelligent services is implicit in the form of unstructured text in various books and journal papers in related fields, which is difficult to capture, use, analyze, and share. In particular, geographical feature knowledge is one of the types of knowledge that needs to be extracted urgently. To solve this problem, in this paper, we propose a method for the rapid extraction of research areas from scientific and technological literature abstracts. Firstly, with the help of a general naming entity identification tool, we propose a method of rapidly annotating place-name entities in administrative divisions. Then, combining the bidirectional long short-term memory conditional random field (BiLSTM-CRF) model with a place-name database covering five levels of administrative divisions in China, the identification, disambiguation, and relationship extraction of place names in different administrative divisions are realized. On this basis, the extraction of research areas is regarded as a two-classification problem, feature vectors such as frequency and location are constructed for the names of the extracted administrative divisions, and the classification model is constructed with the random forest algorithm to rapidly extract research areas. The experimental results show that the recognition accuracy of place names in administrative areas in this study is 92.61% and the recognition accuracy of research areas is 90.31%. The results are superior to those of similar algorithms; thus, the proposed method can accurately and rapidly extract research areas.

1. Introduction

After years of hard work, the field of surveying and mapping geographic information has built a multiscale basic geographic information database system with timely updates, which has played an important role in the construction and application of smart cities.[1,2] In recent

*Corresponding author: e-mail: lwz@ngcc.cn
https://doi.org/10.18494/SAM.2020.3127

ISSN 0914-4935

years, with the gradual development of intelligent city construction, we are required to meet the personalized application needs of users and provide intelligent services with spatial sensing such as the intelligent recommendation of spatial data and the discovery of hotspots to support smart city planning, management, and decision-making research.[3] However, at present, massive data, an explosion of information, and hard-to-find knowledge are phenomena in basic geographic information services, making it difficult to meet the needs of users of geospatial knowledge services and to realize innovation in surveying and mapping science and technology.[4] The main reason for these phenomena is that most of the above-mentioned knowledge exists implicitly in an unstructured form in various books and journal papers in different fields, which makes it difficult to capture, share, and reuse.[5] Therefore, as the foundation of computer understanding of literature, knowledge extraction technology has important research value and broad application prospects.[6]

Journal papers are important carriers of the knowledge of different disciplines in various fields, which condense the excellent research ideas, theories, and achievements of scholars. They are the most cutting-edge, authoritative, and easily accessible knowledge resources in various research fields, including extensive professional core knowledge such as research problems, algorithm models, and other types of knowledge.[7] Facing the demand for geographic information services in the construction of smart cities, where more than 80% of all types of information involved in the development of smart cities are related to geospatial locations, the simulation space support of a smart city is the geospatial framework of the digital city and the geographical framework is the core of a city's efficient operation.[8] Therefore, geospatial knowledge is an important part of constructing a geographic information system for smart city construction. If geospatial knowledge can be extracted from the massive amount of scientific and technological literature, it can provide users with knowledge services such as hotspot discovery, location-based spatial data recommendation, and other services through simple statistical analysis, association rule mining, and so forth.[9] According to different needs, the geospatial knowledge in the literature can be divided into sampling and research areas where scientific research activities are located. Most scholars are dedicated to extracting the names contained in the literature or extracting scientific research events.[10,11] However, the naming entity identification technology cannot determine which place names are related to research areas, and the extraction of scientific research events cannot guarantee that all place names related to research areas can be extracted.

This paper mainly focuses on the extraction of research areas from scientific and technological literature abstracts. First, in view of the inaccuracy of the universal naming entity identification tool, a method of rapid name marking is proposed. By combining the bi-directional long short-term memory conditional random field (BiLSTM-CRF) model with a five-level administrative division place-name database, the extraction, disambiguation, and relationship extraction of the administrative division place names in a document abstract are realized. On the basis of place-name entity recognition, research area identification is abstracted as a two-classification problem, and the random forest classification module is introduced. The classification model is trained by rapidly constructing feature vectors such as the frequency and location of the place names. As a result, the extraction of research areas has high accuracy and practicability.

2. Related Work

The extraction of research areas from literature abstracts mainly refers to the identification of geographic entities that appear in them and determines whether they are the research areas or where the scientific research activities are located. The key extraction technologies mainly include place-name identification and research area extraction.

For the research on text-oriented place-name recognition, the method based on pattern matching has been gradually replaced by supervised machine learning methods, such as the hidden Markov model (HMM) and conditional random field (CRF) models, because of its low recall rate and excessive cost of constructing patterns.[11–13] In recent years, with the rapid development of artificial neural network technology, many scholars have used CRF-nested neural network models, such as IDCNN-CRF, BiLSTM-CRF, and RNN-CRF, to carry out research on the entity recognition of place names. Among them, BiLSTM-CRF is the most popular: BiLSTM can effectively use past and future input features and CRF can help use sentence-level label information. This method can achieve 90% accuracy in certain specific fields.[14–16] However, owing to the lack of research on place-name extraction from scientific and technological literature, there is a lack of directly usable annotation data. In addition, because of the layer-by-layer abstraction of human cognition and the diversification of expressions, the extracted place names have ambiguities, and the disambiguation rules are generally written by linguists.[17,18] Owing to the limited coverage of the rules, the effectiveness of disambiguation by this method is not ideal.[17] With the increasing growth and improvement of the encyclopedia knowledge base, it has become a valuable knowledge source for disambiguation, providing rich expressions, rapid updates, and extensive coverage of background knowledge, making it a new trend in place-name disambiguation.[19,20]

There have been very few studies on research area extraction in the literature. Similar research has mainly focused on the extraction of news events. Early research on the extraction of news events directly used the geographical entities identified in the text as the spatial location of occurrence or directly used the spatial location information attached to the text to assign the location as the place where the news event occurs. Part of the studies considered entity relationships in the extraction process, but they were mainly used in place-name disambiguation, the extraction result was still expressed by a single geographic entity, and the effectiveness of place recognition was unsatisfactory.[20,21] In recent years, many scholars have carried out work on research area extraction from the aspects of dependency syntax analysis, feature construction classification models, and so forth, and have obtained high recognition accuracy. However, the related research corpus is mainly news, Weibo content, and other public opinion data, which does not have good universality. Thus, it is difficult to directly use it with scientific literature data.[22–24]

At present, the main difficulties in identifying geographical entities from scientific research literature are how to rapidly construct annotated data sets and the method of place-name disambiguation. Moreover, there are very few related results on research area extraction. A major challenge is how to construct a classification model based on the semantic characteristics of the literature study area. In addition, it is necessary to combine multiple methods in the research process and incorporate more domain knowledge resources to reduce labor costs and improve research efficiency.

3. Key Technology

3.1 Place-name identification

In addition to the word segmentation characteristics of common Chinese, the hierarchical characteristics of place names and the randomness, diversity, and ambiguity of place names also increase the difficulty in recognizing place-name entities. The multilocation information in the literature adds to these difficulties. The BiLSTM model, which does not rely on dictionaries and features, has strong context memory capabilities. It can solve the problems of unregistered words and ambiguity, while the CRF algorithm can control the address annotation output through a transition probability matrix. Therefore, in this paper, we use the BiLSTM-CRF model to identify the names of administrative divisions in literature abstracts.

3.1.1 Principles of BiLSTM-CRF model

The BiLSTM-CRF model is divided into three layers, as shown in Fig. 1: the presentation BiLSTM, and CRF layers. First, a new training data set is generated by labeling a large number of place names in the document abstract data, and then training is carried out through the Word2vec vector model to form a high-dimensional word vector matrix. The word vector sequence corresponding to each sentence in the training data set is input into the BiLSTM module for feature extraction by looking up the table. Finally, the feature vectors output by the BiLSTM module are sequence-labeled through the CRF module to increase the relevance of text information and improve the accuracy of label prediction.

The identification of place-name entities based on BiLSTM-CRF is a typical sequence labeling problem. The model requires large-scale labeling data support to ensure the accuracy of recognition. However, at present, there is a lack of large-scale marking data for the identification of place names in scientific and technological literature, and the time and labor required for manual marking are very high. Therefore, we propose a rapid labeling method based on an existing word segmentation tool (HanLP) for place-name entities, as shown in Fig. 2.

The labeling method includes five main steps. Because the number of raw data is large, if the labeling process is carried out directly, it will require a lot of time and labor. Therefore,

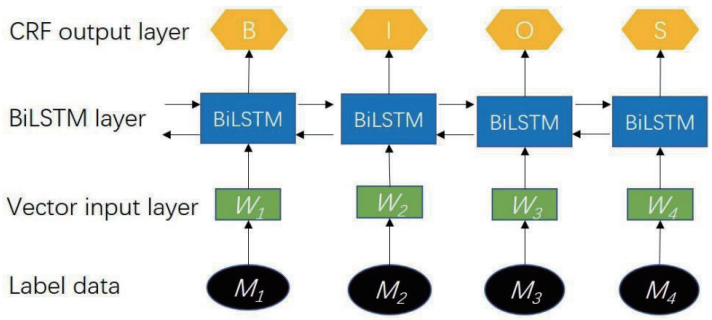

Fig. 1. (Color online) BiLSTM-CRF model structure.

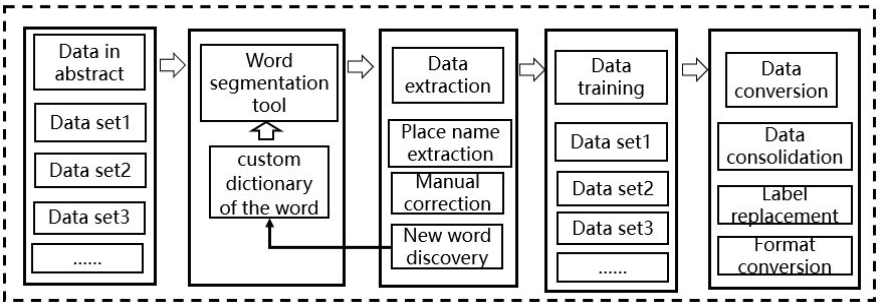

Fig. 2. Flow chart of rapid labeling method for place names in literature abstracts.

the first step is to evenly divide the raw data into multiple sub-data sets. For example, 5000 pieces of data are divided into five data sets, each including 1000 pieces of data. Each data set is segmented in order. After the previous data set is segmented, the user-defined dictionary of the word segmentation tool can be optimized to solve the same problem in the next data set, and it will be easier to process the next data set, reducing the time and labor required to label data. The second step is to use the HanLP word segmentation tool to segment each data set, which is a tool based on words. After importing the abstract, the tool marks words such as place names, organization names, and person names with set labels. This tool can identify more place names by optimizing a custom vocabulary. The third step is to extract the words labeled as place names in the second step to obtain a place-name data set. The abstract is manually read and the place names separated by the word segmentation tool are corrected. If there is an undivided place name, it is manually added to the custom dictionary of the word segmentation tool, and this place name can be recognized when the next data set is segmented. The fourth step is to perform the second and third steps in sequence on each divided data set to obtain the manually corrected place-name data set. Because some problems cannot be solved by optimizing a custom dictionary, we use these corrected data sets as training data and import them into the BiLSTM-CRF model based on characters to train the place-name recognition model, so as to solve the other problems in HanLP word segmentation. The fifth step is to accumulate several parts of the data into a data set, write an algorithm to match the corrected place names with the place names in the abstract, and replace the place names in the abstract with the form "/o place name /ns". Each word in the abstract is a single line. When encountering words that start with "/o" and end with "/ns", "B-LOC" is marked after the first word, the next few words are marked with "I-LOC", and all other words are marked with "O".

3.1.2 Place-name disambiguation and relation extraction based on place-name database of five-level administrative divisions

Using the BiLSTM-CRF model, place names in the literature can be extracted accurately, but because of the nature of the Chinese language, in place-name naming, the ambiguity caused by the same place names will reduce the practicality of the extraction results. In addition, the affiliation relationship between place names is a factor that needs to be considered in the construction of a research area about place-name characteristics. Therefore, we propose a

method for the disambiguation and relationship extraction of place names that is based on a knowledge graph of administrative divisions. The knowledge graph of administrative divisions is the result of the preliminary work of the project team. The knowledge graph contains the main attributes and affiliations of all place names at the five administrative levels of China: province, city, county, township (town), and village. Relevant knowledge service applications have been developed on the basis of this knowledge graph (http://kmap.ckcest.cn/town/tosearch). The method of place-name disambiguation and relation extraction is shown in Fig. 3.

The method mainly includes the following six steps. Step 1: Using the place-name database, accurate, complete, and uniquely matched place names in an abstract are disambiguated. Considering that many place names in an abstract use abbreviated forms, the place names of the place-name database include the full name and the name without the suffix. The matching process first matches the complete place name; if it cannot match the complete place name, it matches the abbreviation, where the abbreviation matching must be unique. Step 2: The set of place names is divided according to the distance between place names in the abstract. If the distance between place names is less than or equal to 1, then these place names may have an affiliation relationship (distance = 0) or a level relationship (distance = 1). These place names are divided into subsets with affiliation or level relations, and then they are matched in the place-name database by semantic similarity calculation, and the matching names are marked. Step 3: After the first two steps of disambiguation, there may be more than two ambiguous items in the place-name database, which can be disambiguated according to the distance between them and the names marked in the first two steps. The shorter the distance, the higher the correct rate of disambiguation. Step 4: If disambiguation cannot be achieved by marking place names, the distance between these ambiguous names can also be calculated in the place-name database, and the place name with the shortest distance can be selected as the correct place name. Step 5: If a single place name cannot be disambiguated through the above four steps, a final disambiguation is performed by considering the administrative division scale of the place name to be matched, and the place name with the highest administrative division level is selected for disambiguation. This is because the geographical location of higher administrative divisions is more likely to be relevant because of the large population and the developed economy. Step 6: For the place name obtained after the disambiguation, the name corresponding to the place-name database can be chosen as its standard name, and its relationship in the place-name database is extracted to provide assistance in the next step of calculating the characteristics of the research area.

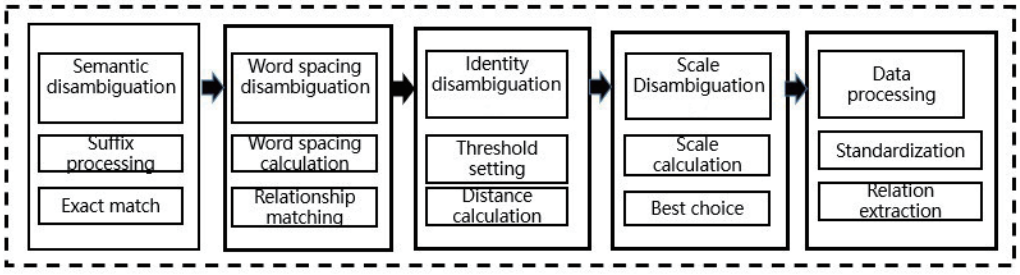

Fig. 3. Flow chart of place-name disambiguation and relation extraction.

3.2 Research area extraction

3.2.1 Random forest

Research area extraction is performed to extract the place of scientific research activity from the place names extracted from the literature abstract. An abstract contains at least one research area. In this paper, the extraction of the research area is regarded as a two-category problem, that is, a research area is divided into two cases: yes or no. At present, there are many classification models, such as naive Bayes, support vector machine, random forest, and classification and regression tree. Among them, the random forest algorithm is easy to implement and has high accuracy. Therefore, we use the random forest model to classify the research area. The classification principle is shown in Fig. 4.

The random forest is a classifier that contains multiple decision trees. It uses n decision trees for classification and a simple voting method to obtain the final classification results, thereby improving the accuracy of classification. In other words, for classification data with an unbalanced distribution, it can also balance the errors generated. In the random forest dichotomy algorithm, the input parameter is the word feature vector. For the research area extraction task, this vector refers to the feature set of each place name in the abstract, including frequency and location characteristics, and other characteristics. Under the premise that the purpose of classification is not clear, the place names in the document abstract can be used to construct feature vectors from multiple dimensions, such as similarity, word frequency, location, distance, and other features. Generally, the more feature dimensions, the higher the classification accuracy, although the time cost also increases. The main task in this paper is to

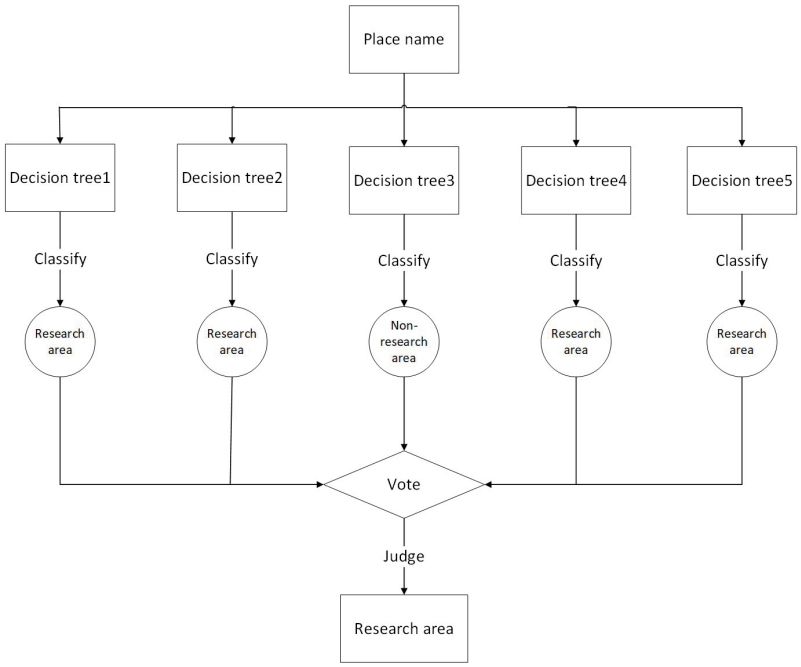

Fig. 4. Schematic diagram of research area extraction based on random forest model.

rapidly extract the research area, which requires both accuracy and efficiency. Therefore, three important characteristics, place name frequency, whether the place name is in the title, and the place name position, are mainly selected for rapid classification.

3.2.2 Classification feature construction

(1) Frequency characteristics of place names

If a place name appears multiple times in the abstract and more frequently than other place names, then this place name is probably the research area of this article. If two place names have an inclusive relationship, the frequency of place names with high administrative divisions is added to that of place names with low divisions. We take "Beijing" and "Xicheng District" as examples. Xicheng District is part of Beijing, so the frequency of "Beijing" is added to that of "Xicheng District". Assuming that the abstract contains three place names a, b, and c, and a is part of b, the frequency calculation formulas of the three place names are as follows:

$$\begin{aligned} f(a) &= \frac{p_a + p_b}{p_a + p_b + p_c} \\ f(b) &= \frac{p_b}{p_a + p_b + p_c}, \\ f(c) &= \frac{p_c}{p_a + p_b + p_c} \end{aligned} \tag{1}$$

where $f(a)$, $f(b)$, and $f(c)$ are the frequencies of place names a, b, and c in the abstract, and p_a, p_b, and p_c represent the numbers of place names a, b, and c, respectively.

To verify the rationality of feature settings, 150 data were extracted for an experiment, in which the frequency of all place names was first calculated, and then they were classified according to whether they were research areas, as shown in Fig. 5, where the ordinate is the frequency of place names. It can be seen that in the literature, the frequency of place names in the research area is generally greater than that in the non-research area. Therefore, the frequency of place names can be used as a characteristic value of the research area.

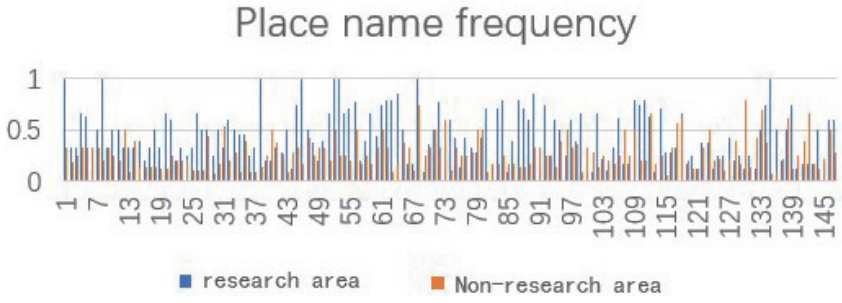

Fig. 5. (Color online) Statistical analysis of frequency characteristics of place names.

(2) Whether the place name is in the title

If an abstract is a condensed summary of the document, then the title can be considered to be a condensed summary of the abstract. The place name mentioned in the title is likely to be the research area. Because the titles of some scientific research documents directly express research at a certain place, whether the place name appears in the title can also be used as a basis for judging whether the place name is a research area. The existence and title of place name a can be expressed by the following formula:

$$H(a)=\begin{cases}0 & \text{Title does not contain place name}\\ 1 & \text{Title contains place name}\end{cases}, \tag{2}$$

where $H(a)$ represents whether the place name is in the title: a value of 0 means that the title does not contain place name a, and a value of 1 means that the title contains place name a. To verify the rationality of the feature setting, 200 pieces of data were extracted for statistical analysis, and the results are shown in Fig. 6. Place names were randomly sampled and calculated. The probability that they existed in the title and were the research area was 55%, and the probability that they existed in the title but were not the research area was 6%, as shown in the figure. It can be seen that the existence of a place name in the title can be used to distinguish whether the place name is a research area, so it can be set as a characteristic value in research area classification.

(3) Location characteristics of place names

In an abstract, the location of the research area also has certain regularity, mostly appearing at the beginning of the abstract and occasionally at the end, so the location characteristics of the place name in the abstract can also be used as a basis for judging whether the place name is the research area. Because the same place name may be distributed throughout the abstract, we only calculate the position where the place name first appears.

The calculation formula for the place-name position is

$$w(a)=\frac{Fa}{Fn}, \tag{3}$$

Fig. 6. (Color online) Statistical analysis of whether place name in title is research area.

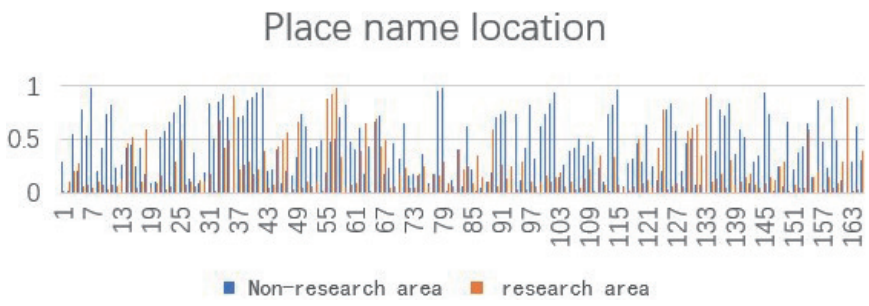

Fig. 7. (Color online) Location characteristics of place names.

where $w(a)$ represents the location feature value of place name a, Fa is the word number where place name a first appears, and Fn is the total number of words in the abstract.

To verify the rationality of the feature setting, 170 pieces of data were extracted for statistical analysis, and the results are shown in Fig. 7. It can be seen that the feature values of place names in the research area are generally small, that is, they are generally at the front of the abstract, and individual feature values are close to 1, that is, near the end of the abstract. Therefore, the feature of a place name can also be used as a feature value of the research area.

4. Experiment and Discussion

4.1 Experimental material

The data used in this research was from the geographic information professional knowledge service platform. At present, the platform has collected more than 10 million articles on surveying and mapping geographic information and related fields (covering the period 1991–2018). We randomly selected 10000 literature abstracts as corpus data. The literature metadata consisted of several fields, such as title, abstract, time, and author. The HanLP tool was used to segment the abstract and select the place names based on the part of the text, and then a second accurate labeling was performed through manual correction. Among the data, 5000 pieces of data were used in an entity recognition experiment on place names. The corpus ratio of the CRF model training set to the test set was about 10:1. The remaining 5000 pieces of data were used in a research area identification experiment and the research area was manually marked. The data volume ratio of the random forest model training set to the test set was about 5:1.

4.2 Experimental

4.2.1 Experimental setup

The configuration of the computer hardware and software and the main parameters of the BiLSTM-CRF and random forest models are shown in Tables 1–3, respectively.

Table 1
Experimental configuration and framework.

Parameter	Value
CPU	Intel Core i7-6700HQ CPU @ 2.60 GHz tetranuclear
GPU	Nvidia GeForce GTX 950M 4G
Operating system	Windows 10 64 bits
Programming language	Python
Deep learning framework	TensorFlow

Table 2
Parameters of BiLSTM-CRF model.

Parameter	Value
batch_size	64
epoch	20
hidden_dim	300
dropout	0.5
optimizer	Adam
embedding_dim	300
learning rate	0.001

Table 3
Parameters of random forest model.

Parameter	Value
max_depth	None
max_features	Auto
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0.0
n_estimators	15
n_jobs	None
verbose	0

4.2.2 Model evaluation indicators

By comparing the indicators, the effectiveness of the model is evaluated. We set the recall rate (Recall), precision (Precision), and $F1$ value to evaluate the named entity recognition model. The main evaluation indicators are these three indicators and accuracy. For the two-class model, it is inaccurate to judge the accuracy of the research area only, so two indicators, the average (macro avg) and the weighted average (weighted avg), are added. The average value index is used when the sample ratio of the research area to the non-research area is about 1:1, and the weighted average value index is used when the ratio is out of balance.

The formula for calculating the recall rate R is

$$R = TA / FB, \tag{4}$$

where TA is the number of toponyms correctly identified as the study area and FB is the total number of toponyms of the study area.

The formula for calculating the accuracy rate P is

$$P = TA / FA, \tag{5}$$

where FA is the number of samples.

The formula for calculating the $F1$ value is

$$F1 = 2 \times P \times R / (P + R). \tag{6}$$

5. Experimental Results and Analysis

5.1 Place-name recognition results

The batch size of the model represents the amount of data read in the model training network, and the epoch represents the number of iterations. The two parameters mainly affect the time cost and performance of the model training. First, we select 1000 pieces of training data to tune the two parameters. The 1000 pieces of data are independent of the 10000 pieces of data used for modeling mentioned in the next paragraph. When the batch size is 64 and the epoch is 20, the model can obtain the local optimal solution with the lowest time cost. The number of nodes (H) and the learning rate (LR) are the two training parameters that mainly affect the training accuracy of the model. To obtain the best experimental results, we set the batch size to 64 and the epoch to 20, and select 1000 pieces of training data for parameter-tuning experiments. The training accuracy for four different parameter configurations is shown in Table 4.

It can be seen that when H is 300 and LR is 0.001, the accuracy of place-name recognition of the model is the highest. To verify the superiority of the proposed method, in the next experiment, the optimal model parameters are used, 5000 training data sets are selected, and the CRF, BiLSTM, and LSTM-CRF models are used for comparison with the BiLSTM-CRF model. The performance characteristics of the four models are shown in Table 5.

According to Table 5, the accuracy indicators of the CRF and BiLSTM models are relatively close. The accuracy indicators of the LSTM-CRF model are better than those of the first two single models, but because the LSTM network in the sequence-labeling model can only extract context features, the extraction model does not achieve the best results. Compared with the other three model methods, the BiLSTM-CRF model has higher precision and recall rates, which shows that the method based on the BiLSTM-CRF model is superior to the other methods.

Table 4
Recognition performance of model for different training parameters.

Parameters	Precision (%)	Recall (%)	$F1$ (%)
$H = 200$, $LR = 0.001$	63.23	71.57	67.10
$H = 200$, $LR = 0.003$	72.62	67.83	70.13
$H = 300$, $LR = 0.001$	75.35	74.77	75.06
$H = 300$, $LR = 0.003$	71.28	69.43	70.34

Table 5
Experimental results for test set obtained with different models.

Model	Precision (%)	Recall (%)	$F1$ (%)
CRF	83.74	83.96	83.63
BiLSTM	87.31	86.64	85.65
LSTM-CRF	89.99	88.12	89.64
BiLSTM-CRF	92.38	92.45	92.61

5.2 Extraction results for the research area

The random forest model has many parameters, with three parameters, n_estimators, max_depth, and max_features, having an important impact on the accuracy of the model. n_estimators represents the maximum number of iterations of the learner. Generally, if the value is very small, underfitting may occur. If it is very large, the cost will increase and the performance will not significantly increase. By selecting 1000 training data for many experiments, we find that setting this value to 15 gives the best performance. Max_depth and max_features respectively represent the maximum number of features considered when constructing the optimal model of the decision tree and the maximum depth of the decision tree. On this basis, only three features are selected in the next experiment. Therefore, the maximum value of both parameters is selected without limitation. Max_depth and max_features are respectively set to None and Auto. To verify the superiority of the proposed method, a training set of 5000 data is selected under the above-mentioned optimal model parameter configuration, with the naive Bayes model, K-proximity method model, decision tree model, and SVC used for comparison with the random forest model. The results for the five models are given in Table 6.

The random forest algorithm has the best classification performance. Among the five algorithms, the naive Bayes algorithm is the simplest. It is generally used in text classification, but it does not perform well for the research area/non-research area dichotomy problem in this article. The algorithm of the K-nearest method of the SVC model has better classification performance than the naive Bayes algorithm and also requires fewer samples than the SVC model to achieve the same accuracy. For a given sample size, the K-proximity method gives superior results to the SVC model. The decision tree model achieves good results for large data sources in a relatively short time. The size of the training set in this experiment is 5000, so the accuracy of the algorithm is higher than those of the SVC and K-proximity methods. However, because a large amount of training data is prone to noise, the decision tree is prone to using noisy data as the separation standard, which often leads to overfitting. The random forest algorithm uses the voting mechanism of multiple decision trees to reduce the overfitting problem of the decision tree, and the classification performance result is better than that of the decision tree model.

Table 6
Results for different models.

Algorithm	Precision (%)	Recall (%)	*F*1 (%)
Naive Bayes	77.31	76.64	75.65
SVC	82.11	82.24	81.77
K-proximity	83.99	84.12	83.64
Decision tree	86.15	87.20	86.64
Random forest	91.38	90.55	90.31

6. Conclusion

Aiming to solve the problem of the overflow of information and the lack of knowledge faced by intelligent geographic services with spatial sensing, we propose a method of extracting knowledge on the location of research areas from scientific and technological literature. Place-name recognition is an important basic task in extracting research areas. Therefore, we carried out the first ever recognition of place names using the BiLSTM-CRF model. The method of NER combined with manual correction can ensure the accuracy of place-name recognition and greatly reduce labor costs. With the help of a five-level place-name knowledge map to disambiguate the recognized place names and extract relations, we can further improve the practicability of place-name recognition. On this basis, we construct the characteristics of the frequency and location of place names in the research area using the random forest classification algorithm, which rapidly and accurately extracts the study area of the literature abstract, and the data with greater accuracy is better than similar algorithms. Although the place names of the research areas extracted in this paper are those of administrative districts, there are also natural geographical entities such as water systems and mountain ranges in the scientific and technological literature. Therefore, in the next step of this research, a large amount of labeling data needs to be added with the help of a larger geographical knowledge atlas to realize the recognition, disambiguation, and relation extraction of place names. In addition, it is necessary to construct more comprehensive and easy-to-implement classification features to further improve the accuracy of research area identification.

Acknowledgments

This study was funded by Beijing Key Laboratory of Urban Spatial Information Engineering (No. 2020202) and Geographic Information Professional Knowledge Service System (No. CKCEST-2020-1-5).

References

1 Q. Gong: Inf. Technol. **2** (2017) 25. https://doi.org/10.13274/j.cnki.hdzj.2017.02.006
2 T. H. M. D. Oliveira and M. Painho: Proc. 2015 10th Iberian Conf. Information Systems and Technologies (IEEE, 2015) 1–4. https://doi.org/10.1109/CISTI.2015.7170469
3 D. Li, J. Shan, Z. Shao, X. Zhou, and Y. Yao: J. Geospatial Inf. Sci. **16** (2013) 1. https://doi.org/10.1080/10095020.2013.772803
4 J. S. Ning: Mapping Geoinf. Technol. **14** (2016) 2. https://doi.org/10.3969/j.issn.1672-4623.2016.02.001
5 C. Wang, X. Ma, J. Chen, and J. Chen: Comput. Geosci. **112** (2018) 112. https://doi.org/10.1016/j.cageo.2017.12.007
6 M. Balakrishna, S. Werner, and M. Tatu: Proc. IEEE Tenth Int. Conf. Semantic Computing (IEEE, 2016) 390–391. https://doi.org/10.1109/ICSC.2016.30
7 X. Ma, Z. Xuan, and J. Wu: Proc. 2010 Int. Conf. E-Product E-Service and E-Entertainment (IEEE, 2010) 1–4. https://doi.org/10.1109/ICEEE.2010.5660824
8 S. Roche: Prog. Human Geogr. **38** (2014) 315. https://doi.org/10.1177/0309132513517365
9 N. Freire, J. Borbinha, and P. Calado: Proc. The Semantic Web: Research and Applications (ESWC, 2012) 718–732.
10 N. Freire, J. L. Borbinha, and P. Calado: Proc. 2011 Joint Int. Conf. Digital Libraries (JCDL, 2011) 13–17. https://doi.org/10.1145/1998076.1998140

11 Y. Wei, H. U. Danlu, and L. I. Xiang: J. Geomatics Sci. Technol. **32** (2016) 6. https://doi.org/10.3969/j.issn.1673-6338.2016.01.019
12 X. Li, X. Lv, and K. Liu: Proc. CCF Int. Conf. Natural Language Processing and Chinese Computing (NLPCC, 2014) 379–391. https://doi.org/10.1007/978-3-662-45924-9_34
13 K. Nongmeikapam, T. Shangkhunem, and N. M. Chanu: Proc. 2011 2nd National Conf. Emerging Trends and Applications in Computer Science (IEEE, 2011) 1–6. https://doi.org/10.1109/NCETACS.2011.5751390
14 Y. Jia and X. Xu: Chin. Comput. Commun. **9** (2019) 41. https://doi.org/10.1109/ICSESS.2018.8663820
15 D. C. Wintaka, M. A. Bijaksana, and I. Assor: Procedia Comput. Ence. **157** (2019) 221. https://doi.org/10.1016/j.procs.2019.08.161
16 E. Ouyang, Y. Li, and J. Ling: Proc. China Conf. Knowledge Graph and Semantic Computing (CCKS, 2017).
17 Y. Ju, B. Adams, and K. Janowicz: Springer Int. Publishing **10024** (2016) 353. https://doi.org/10.1007/978-3-319-49004-5_23
18 Y. Hu, K. Janowicz, and S. Prasad: Proc. 8th Workshop on Geographic Information Retrieval (2014) 1. https://doi.org/10.1145/2675354.2675356
19 J. Lehmann, R. Isele, and M. Jakob: Semantic Web **6** (2014) 2. https://doi.org/10.3233/SW-140134
20 D. Buscaldi and P. Rosso: Proc. 5th ACM Workshop on Geographic Information Retrieval (2008) 19. https://doi.org/10.1145/1460007.1460011
21 Z. Ying, H. Yang, and Y. Feng: Proc. Natural Language Understanding and Intelligent Application (NLPCC, 2016) 275. https://doi.org/10.1007/978-3-319-50496-4_23
22 S. Unankard, X. Li, and M. A. Sharaf: World Wide Web-Internet Web Inf. Syst. **18** (2015) 1393. https://doi.org/10.1007/s11280-014-0291-3
23 S. Kinsella, V. Murdock, and N. O. Hare: Proc. 3rd Int. Workshop on Search and Mining User-generated Contents (SMUC, 2011) 61-68. https://doi.org/10.1145/2065023.2065039
24 Y. Fan: Research on Event Location Extraction of News[D] (Harbin Institute of Technology, Harbin, 2018).

About the Authors

Chuan Yin received his B.E. degree from Sichuan Normal University, China, in 2009 and his M.S. and Ph.D. degrees from Capital Normal University, China, in 2012 and 2015, respectively. From 2015 to 2018, he did postdoctoral research at National Geomatics Centre of China. Since 2018, he has been a lecturer at Beijing University of Civil Engineering and Architecture. His research interests are in semantic webs and knowledge maps.
(yinchuan@bucea.edu.cn)

Wanzeng Liu received his B.S. degree in mine surveying, his M.S. degree in management science and engineering, and his Ph.D. degree in cartography and geographic information systems from China University of Mining and Technology in 1992, 2000, and 2005, respectively. From 1992 to 1997, he was an assistant engineer at Yongxia Mining Area Construction Management Commission, China. From 1997 to 2002, he was an engineer at Yongcheng Coal Power Group, China. Since 2006, he has been a senior engineer at National Geomatics Centre of China. His research interests are in emergency mapping, spatial knowledge services, and geographic information systems.
(lwz@ngcc.cn)

Duoduo Yin has been a graduate student in geographic information systems at College of Resource Environment and Tourism, Capital Normal University, since 2019. Her research has been concerned with the methods and applications of GIS in natural language processing and deep learning. She is interested in solving problems in Chinese natural language processing. (yinduo1996@163.com)

Xi Zhai received his B.S. degree in engineering of surveying and mapping from ShanDong Agricultural University, China, in 2006 and his M.S. and Ph.D. degrees in cartography and geographical information engineering from Wuhan University, China, in 2012 and 2017, respectively. Since 2017, he has been an engineer at National Geomatics Centre of China. His research interests include emergency mapping, sensor webs, and spatial knowledge services. (zhaixi@ngcc.cn)

Kexin Liu received her B.S. degree in geographic information systems in 2020 from Beijing University of Civil Engineering and Architecture, China. She is currently studying for a second degree in business administration from Beijing University of Civil Engineering and Architecture. (lkx0104xzg@163.com)

Changfeng Jing received his B.S. degree in surveying and mapping from the China University of Petroleum in 2002 and his Ph.D. degree in geography from Zhejiang University in 2008. Since 2008, he has been working at Beijing University of Civil Engineering and Architecture (BUCEA). Since 2015, he has been an associate professor with School of Geomatics and Urban Spatial Informatics, BUCEA. He is the author of two books and more than 30 articles, and has patented more than 10 inventions. His research interests include urban spatiotemporal analysis, geostatistics, urban Internet of Things, and urban planning management. (jingcf@bucea.edu.cn)

He Huang received his B.S. degree in surveying engineering from Wuhan Technical University of Surveying and Mapping, China, in 2000 and his M.S. and Ph.D. degrees from SungKyunKwan University, Korea, in 2004 and 2010, respectively. Since 2010, he has been a lecturer and an associate professor at Beijing University of Civil Engineering and Architecture, China. His research interests are in high-precision intelligent driving navigation maps and visual navigation and positioning. (huanghe@bucea.edu.cn)