# Designing of a Geographical Entity Annotation System Using the BiLSTM+CRF+AGG Model

Aiping Xu , Guangming Ling , Chao Wang *

Posted Date: 26 September 2023

doi: 10.20944/preprints202309.1773.v1

Keywords: geospatial entities; annotation system; deep learning; BiLSTM+CRF+AGG model; active Learning; human-assisted

*Article*

# Designing of a Geographical Entity Annotation System Using the BiLSTM+CRF+AGG Model

**Aiping Xu [1,2,‡], Guangming Ling [2,‡] and Chao Wang [3,*]**

1 School of Computer Science, Wuhan Donghu University, China
2 School of Computer Science, Wuhan University, China
3 The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, China
* Correspondence: c.wang@whu.edu.cn; Tel.: +86-18986074931 (Chao Wang)
‡ These authors contributed equally to this work.

**Abstract:** When analyzing user geospatial information through deep learning methods, it is typically necessary to annotate existing geospatial data. Currently, manual annotation methods are commonly employed, suffering from issues related to low efficiency and accuracy. This design is based on the TensorFlow deep learning framework and first realizes the BiLSTM+CRF+AGG deep learning model. Among the models, AGG is the aggregation layer, which is introduced to solve the problem of solid particle size equilibrium. Secondly, based on the characteristics of original data and professional data, an automatic annotation algorithm is proposed. The algorithm first preprocesses the acquired original data set and professional data set, The most valuable unlabeled data set which can make the training model converge quickly is selected from the original data set as the target data set. Sort the target data set based on preset rules and set annotation parameters for the sorted target data set. Based on the set annotation parameters, the corpus is synthesized and used as the annotation result. Thirdly, based on the active learning strategy, a manual annotation auxiliary scheme is proposed, and an Excel generation module that is convenient for manual annotation correction is designed and developed to further improve the efficiency and quality of annotation through iterative processing. The combination of the BiLSTM+CRF+AGG deep learning model and high-quality annotation can accurately identify non-standard, incomplete, or even incorrect geographical information entities. This annotation method has found successful practical application within the context of the research project and has been granted an invention patent.

**Keywords:** Geospatial Entities; annotation system; deep learning; BiLSTM+CRF+AGG Model; active learning; Human-Assisted

## 1. Introduction

Address information, as an indispensable foundation of geographical and social public information, plays a crucial role in national governance, economic development, cultural heritage, and other areas [1,2]. In 2019, the Ministry of Natural Resources of China issued the "Technical Outline for the Construction of Smart City Spatiotemporal Big Data Platforms," explicitly required to realize efficient, accurate, and practical address-matching service technology to establish the orderly organization and correlation of spatiotemporal data and non-spatiotemporal data in smart city [3].

This research takes the project of the National Key Research and Development Program's key project, "Research on Key Technologies for Social Safety Information Services and Big Data Applications." as the background. The main task of this subproject is to automatically extract the time information, location information, and specific parameter information required for three-dimensional scene construction from provided intelligence data through a semantic analysis model. The goal is to achieve the reconstruction of real-life safety accident scenes and dynamic simulations of accidents.

Entity tagging refers to adding metadata to specific entities in text such as people's names, place names, organization names, etc. for easy understanding and processing by machine learning

 doi:10.20944/preprints202309.1773.v1



models [4]. For example, here's a news story about U.S. President Barack Obama: "美国总统奥巴马在白宫与记者见面，讨论了美国与伊朗的核协议问题.奥巴马表示，美国将继续支持伊朗核协议，以确保中东地区的和平与稳定.他还强调了美国对全球气候变化问题的关切，并承诺将采取更多措施应对这一挑战 (President Obama met with reporters at the White House and discussed the issue of the nuclear agreement between the United States and Iran. Obama stated that the United States will continue to support the Iran nuclear agreement to ensure peace and stability in the Middle East. He also emphasized America's concern about global climate change and pledged to take further measures to address this challenge)." In this example, we can make the following entity annotation: character:奥巴马 (President Barack Obama)，记者 (reporter). Location:白宫 (White House)，中东地区 (Middle East region). Organization:美国 (United States)，伊朗 (Iran). Incident: 美国与伊朗的核协议问题 (the nuclear agreement between the United States and Iran)，全球气候变化问题 (global climate change issue).

With these entity annotations, machine learning models can better understand the text content and extract relevant information when needed. With the rapid development of information technology, various text data containing geographical information have emerged in abundance. Especially, data collection channels exhibit diversity and randomness due to historical reasons and work scenarios. This is primarily manifested in two aspects: "lack of standardization in geographical information" and "common occurrence of typographical errors." These factors pose challenges to precise and efficient analysis. On the other hand, recognizing geographical information requires high-quality annotated data to achieve ideal results, but generating high-quality annotated data entails significant manpower costs. Given these issues. This paper presents a deep learning-based geospatial information annotation method and its design and implementation to address, or at least partially address, the technical problems of low efficiency and accuracy in existing methods.

The research contributions of this study can be summarized as follows:

1. An aggregation layer suitable for fine-grained annotation strategy is added and a unique geospatial information analysis model (BiLSTM+CRF+AGG) is formed based on the BiLSTM+CRF deep neural network architecture. After improvement and optimization, the model gives the system have strong fault tolerance ability, and realizes more accurate geographical location information recognition on the non-standard, incomplete text information with wrong characters.
2. The automatic annotation system is realized, which includes data crawling, data cleaning, data integration, data sorting, multi-mode matching, data segmentation, annotation rule definition, and annotation generation functionalities. Additionally, a quality assessment module has been designed and implemented to ensure the high quality of the annotations, significantly enhancing the efficiency of the annotation system.
3. A manual annotation assistance scheme is provided. The Excel generation module is designed and developed for manual correction of annotations. Based on the active learning strategy, the iterative mechanism is further introduced to solve the problem of quantity and quality of data. The system makes the tedious task of tagging lively and interesting.

This paper consists of five sections. The first section introduces this paper's research significance, purpose, content, and results. The second section presents the research work related to this paper and lays the foundation for the research method of this paper. The third section offers the research methods of this paper, including building the BiLSTM+CRF+AGG deep model, automatic annotation process and algorithms, and iterative processing based on active learning. The fourth section is the experimental and practical applications. The fifth section is the conclusion of the full text and further research.

## 2. Related Work

The analysis methods of user geographic information are mainly divided into rule-based methods and statistic-based methods. Rule-based methods are intuitive, natural, easy for human comprehension,

 doi:10.20944/preprints202309.1773.v1



and extensible. However, rules writing depends on specific language knowledge and domain knowledge, so the rules are complicated, difficult to cover all patterns, and the portability is poor. Statistical methods, on the other hand, require less linguistic and domain knowledge and exhibit strong portability, but necessitate manual annotation of corpora and the selection of appropriate statistical learning models and parameters. He et al. [5] introduced a Chinese toponym recognition method based on composite features, combining toponym element features, part-of-speech features, and syntactic features, using CRF (Conditional Random Fields) for Chinese toponym training and recognition. With the advancement of deep learning, Graves et al. [6]combined forward and backward LSTM (Long Short-Term Memory) [7] models to propose the BiLSTM (Bi-directional Long Short-Term Memory) model, effectively utilizing past and future input information. Lample et al. [8] further incorporated CRF to enable sentence-level label information, forming the BiLSTM-CRF model. Huang et al. [9] and Dong et al. [10] proposed and strongly promoted the widespread application of the BiLSTM-CRF deep learning model in NER (Named Entity Recognition) tasks while maintaining robustness with minimal dependency on character vectors. Shen et al. [11] demonstrated the superiority of LSTM decoders over CRF decoders, especially in scenarios with a large number of entity types, where LSTM decoders exhibited faster training speeds. Utilizing character embeddings, the BiLSTM+CRF model eliminates the need for manual feature engineering and effectively addresses the "Feature Engineering" challenge. However, it requires a substantial amount of manually annotated data, resulting in not only low efficiency but also high labor costs.

The cost of manual annotation restricts the development of deep learning to some extent, so the research of automatic annotation algorithms is becoming more and more important. Automatic image annotation techniques have been developed relatively early [12–15]. In the text domain, Huang et al. [16] proposed an automatic annotation method that combines multiple features. Qiu et al. [17] and Chou et al. [18] applied automatic annotation methods to geospatial entity annotation in microblog data. Fu et al. [19] conducted extensive experiments to demonstrate that the BiLSTM-CRF model outperforms CRF, RNN, and BiLSTM models in part-of-speech tagging even without incorporating any manual features. Wang et al.[20] proposed a corpus construction method of geographic entity relations based on annotation technology to realize automatic corpus construction. Zhu et al.[21] conducted a systematic and in-depth study on the evaluation of sample confidence, and proposed an evaluation object extraction method combining active learning and automatic annotation. Schulz et al. [22] conducted in-depth research on dialogue-level sequence annotation tasks, introducing a suggestion model based on BiLSTM-CRF to provide annotation recommendations for domain experts. This model positively impacts annotation speed and performance without introducing significant bias. These studies have actively and effectively explored automatic annotation in various tasks, offering important references for addressing the technological bottlenecks in deep learning.

In this study, based on the BiLSTM+CRF deep neural network model [23], an aggregation layer suitable for fine-grained annotation strategies is added to form a unique geographic information analysis model (BiLSTM+CRF+AGG), which makes the system have strong fault tolerance and annotation efficiency. At the same time, the research also designed and implemented the manual annotation auxiliary module and the detection module to evaluate the annotation quality, and further ensure the annotation quality through the iterative mechanism of active learning.

## 3. Methodology

The automatic annotation system comprises a series of operations, including data crawling, data cleaning, data integration, sorting, multi-mode matching, data segmentation, Excel export (For manual correction convenience), annotation rule definition, annotation generation, and post-annotation processing. Firstly, the automatic annotation algorithm is used to generate the initial training set, and then the BiLSTM+CRF+AGG model is trained with this training set. At this time, only part of the data is used, and the remaining unlabeled data is labeled by the active learning method. In other words, some samples are selected according to the selection strategy, marked with high quality by manual

 doi:10.20944/preprints202309.1773.v1



labeling, and then merged into the training set for training again. This is repeated until active learning ceases.

### *3.1. Build BiLSTM+CRF+AGG Deep Model*

The study found that geographical location granularity is crucial. If the granularity is too coarse, it can lead to a decrease in the model's recognition accuracy. Conversely, if the granularity is too fine, it not only increases the system's computational load but also results in fragmented recognition outcomes. For example, "Wuhan City in Hubei Province" should be whole with type loc, but two entities are identified: Hubei Province and Wuhan City. To improve the recognition accuracy, provide the generalization ability of the network model, and solve the contradiction between the size of the training set and the training cost, the granularity is as small as possible. However, both in the automatic annotation stage and in the prediction stage, fine-grained entities need to be aggregated, so the AGG layer is introduced. Therefore, an improved BiLSTM+CRF+AGG model with an aggregation layer (AGG) is proposed in this study. As shown in Figure 1, it largely solves the entity granularity equalization problem.

In Figure 1, BiLSTM+CRF [23] is a combination model of the BiLSTM (Bi-directional Long Short-Term Memory) layer and CRF (Conditional Random Fields) layer. First, each word in the sentence is represented as a vector and input into the feature embeddings layer, which includes embeddings of words and embeddings of characters. Character embeddings are randomly initialized, and word embeddings are imported from a pre-trained word embedding file. All embeddings will be fine-tuned during training. The input to the BiLSTM model is these embedded features, and the output is the predicted label score of the words in the sentence. All the scores predicted by the BiLSTM layer are then entered into the CRF layer. The label sequence with the highest prediction score in the CRF layer is selected as the best answer. AGG is an aggregation layer, which integrates fine-grained fragmentary entities into a complete entity according to the spatial distribution law of addresses to solve the problem of entity granularity balance.

This approach leverages the BiLSTM+CRF neural network model and develops an automatic annotation system to annotate customer information containing a significant amount of geographical data. Subsequently, it employs the trained high-quality deep neural network model to achieve precise geospatial information recognition.

Based on the TensorFlow deep learning framework, the BiLSTM+CRF+AGG network model is implemented. Among them, AGG is the aggregation layer, which is introduced to solve the problem of entity granularity balance in the NER task. Based on the features of original data and professional data, and integrated with an automatic annotation algorithm, an iterative annotation algorithm based on active learning is formed. Combined with Excel-assisted design technology, manual optimization of annotation data becomes easy and efficient. The combination of high-quality annotation and BiLSTM+CRF+AGG deep learning model can accurately identify non-standard, incomplete, or even incorrect geographical information entities to achieve ideal practical application results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1



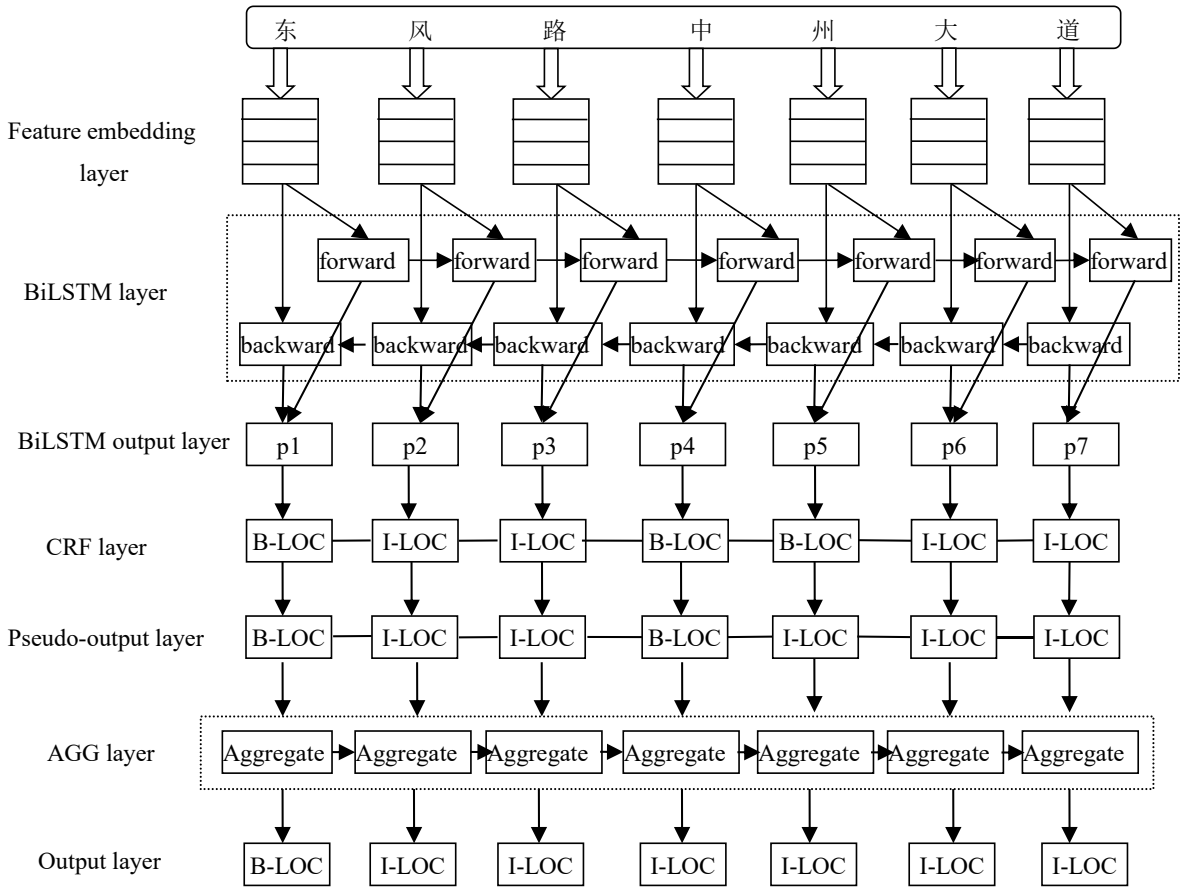


**Figure 1.** The BiLSTM+CRF+AGG deep learning model structure.

### *3.2. Automatic Annotation Process and Algorithms*

The automatic annotation method in this research is based on deep learning models implemented in TensorFlow [24,25]. Besides model design, the corpus is also a critical aspect. This approach requires a substantial amount of community names, which are professional data. In practice, 5341 valuable community names were obtained through web crawling and preprocessing [26].

Some examples of data from the original dataset *UserInfSet* are shown in Table 1. To protect user privacy, Chinese characters are represented by "@" and numbers are represented by "*".

**Table 1.** The examples of the original dataset.

| Examples |
|---|
| 姬@@ 1********2　航海路紫荆山路金锣湾万福园*-*-1**2 |
| 刘@@ 1********3　十八里河新居**号楼*单元1*** |
| 袁@ 1*********1　莲花街翰林国际城桃李园*号楼*单元9** |
| 贾@@ 1********6 1********6 1********6 怡馨家园1-*-1**1 |
| 李@@ 1********0　桐柏南路帝湖花园1**东*单元 |

Some examples of data from the professional dataset *PlotSet* are shown in Table 2.

**Table 2.** The examples of the professional dataset.

| Examples |
|---|
| 河南省水利电力对外公司家属院 |
| 中国人民银行郑州培训学院家属院 |
| 中化地质矿山总局河南地质勘查院 |
| 河南工业贸易职业学院家属院南区 |
| 郑州市市直机关事务管理局家属院 |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1



This method adopts the BIO annotation scheme and according to the data characteristics, five types of labels are defined: PER, TEL, LOC, PLT, and INF, which are used to identify the user name, contact number, road information, cell name, and cell details respectively.

The automatic annotation process is depicted in Figure 2 [27].

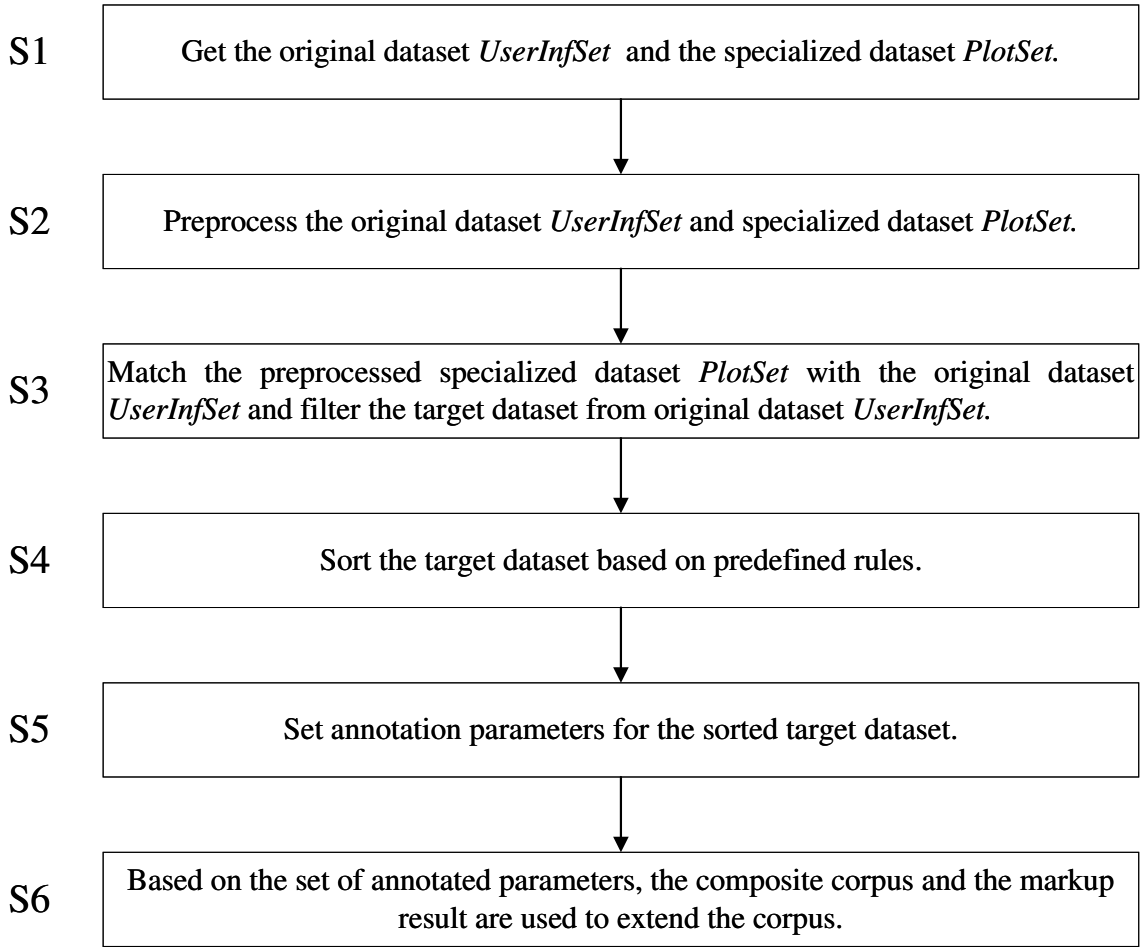


**Figure 2.** The automatic annotation process.

Step S1: Get the original dataset $UserInfSet$ and the specialized dataset $PlotSet$;

Step S2: Preprocess the original datasets $UserInfSet$ and the specialized dataset $PlotSet$.

Step S3: Match the preprocessed specialized dataset $PlotSet$ with the original dataset $UserInfSet$ and filter the target dataset from the original dataset $UserInfSet$.

Step S3.1: Read one piece of raw data $UserInf$ from the original $UserInfSet$.

Step S3.2: Match the raw data $UserInf$ with the professional data in the Hash table $PlotsHashTable$. Obtain the number of communities $PlotCount$ contained in $UserInf$ and match information $MatchInf$. The matching information includes the matching start and end positions.

Step S3.3: If $PlotCount$ is 0, discard the raw data and proceed to Step S3.1. If $PlotCount$ is 1, proceed to Step S3.4. If $PlotCount$ is greater than 1, determine spatial relationships based on the start and end positions, merge according to the spatial relationships, and check if the number of communities after merging is 1. If it is 1, proceed to Step S3.4; otherwise, return to Step S3.1.

Step S3.4: Record the selected raw data $UserInf$ and matching information $MatchInf$ in the target dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1



Step S3.5: Check if all data in the original $UserInfSet$ have been processed. If finished, use the results obtained in Step S3.4 as the target dataset. Otherwise, return to Step S3.1 to process the next data.

Step S4: Sort the target dataset based on predefined rules. The sorted target data set is segmented according to the preset identification, and N Excel files are obtained, where N is an integer greater than 1. By manual adjustment, N Excel files are adjusted to obtain high-quality annotation data.

Step S5: Set annotation parameters for the sorted target dataset. By setting these parameters obtain the predefined annotation standard BIO.

Step S6: Generate a corpus based on the set annotation parameters and treat it as the annotation result, then extend the synthesized corpus.

The central algorithm for implementing the annotation process described above can be found in Algorithm 1.

**Algorithm 1** The automatic annotation

**Require:** The original dataset $UserInfSet$ and the specialized dataset $PlotSet$
**Ensure:** higher-quality annotations

1: Obtain the original data $UserInfSet$ and professional data $PlotSet$, and preprocess them appropriately to ensure data uniqueness and validity.
2: Initialize the resources required for fast matching and construct a hash table called $PlotsHashTable$ to store professional data.
3: **while** $UserInfSet$ has not been processed completely **do**
4: Read one piece of raw data $UserInf$.
5: Iterate through $PlotsHashTable$ to compute matching information $MatchInf$. The information records the matching start and end position and name information in the cell. The $PlotCount$ represents the number of communities, i.e., the number of communities contained within $UserInf$. The cell names in the actual record are not all the same name but are often combined, so $PlotCount$ may take 2,3, or even 4.
6: **if** $PlotCount$ is 0 **then**
7: Discard the data and proceed to line 4.
8: **else if** $PlotCount$ is 1 **then**
9: Proceed to line 18.
10: **else if** $PlotCount$ is greater than 1 **then**
11: Merge based on spatial relationships. The cell is regarded as a line segment in geometric space. This results in more accurate community information.
12: **end if**
13: **if** $PlotCount$ becomes 1 **then**
14: Proceed to line 18.
15: **else**
16: Proceed to line 4.
17: **end if**
18: Record $UserInf$ and $MatchInf$ together in the high-quality dataset $SuperiorSet$.
19: **end while**
20: Randomly sort $SuperiorSet$.
21: Split $SuperiorSet$ based on the predefined list, resulting in N Excel files. These Excel files are designed to be easily adjusted manually to obtain higher-quality annotations.
22: Set annotation parameters.
23: Using the annotation parameters and Excel files, generate the corpus.
24: Extend the corpus.

Finally, after manually adjusting the N Excel files to obtain high-quality annotation data, evaluate the quality of the obtained annotations.

### *3.3. Iterative Processing Based on Active Learning*

Deep learning requires a sufficient amount of high-quality annotated data to achieve ideal results. However, high-quality annotated data requires a large labor cost and is bound to be limited. Therefore, this research adopts an active learning strategy [25]. Active learning is a method to select the most convergent training model from a large amount of information data according to specific selection

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1



rules [28]. It identifies the unlabeled data with the maximum and most valuable information content as a query sample set, enabling fast and accurate matching, and filtering of high-quality, reliable datasets for subsequent annotation work. The processing flow based on active learning is depicted in Figure 3.

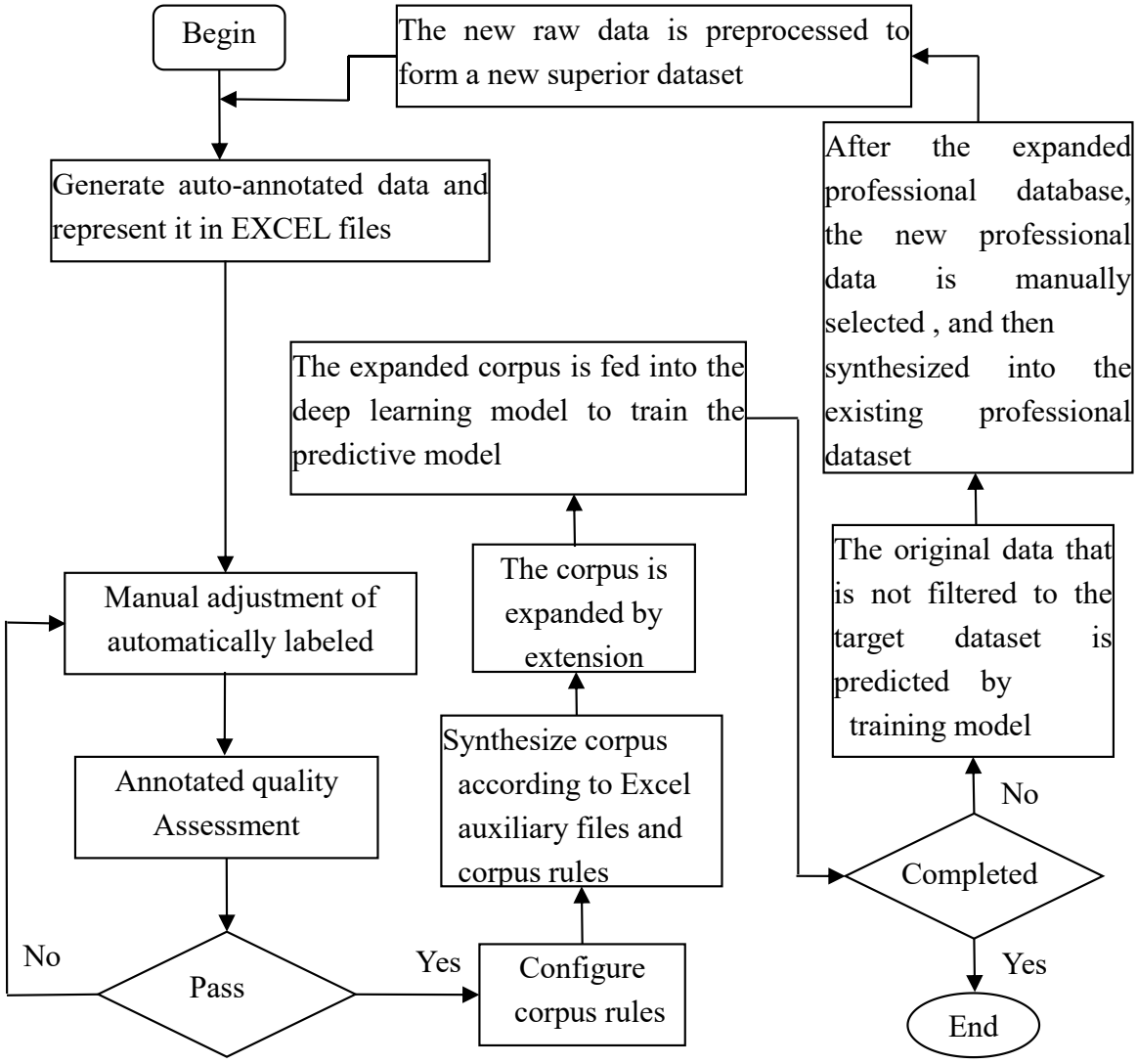


**Figure 3.** The Evaluation post-processing process based on active learning.

In Figure 3, firstly, the auto-annotated data is generated and represented in Excel files, and then the auto-annotated corpus is manually adjusted and the annotation quality is evaluated. The quality of annotations can be assessed by the coverage rate of key annotations. If it does not pass the evaluation, the corpus will be adjusted manually until it passes the evaluation. After passing the quality assessment, the corpus rules are configured and the corpus is synthesized according to the rules. The corpus is expanded by extension modules. The expanded corpus is fed into the deep learning model to train the predictive model. The newly generated specialized data is used to determine whether iteration is necessary, and if iteration is required, the original data that is not filtered to the target dataset is predicted by the trained model. After the expanded professional database, the new professional data is manually selected and then synthesized into the existing professional data. The new raw data is preprocessed to form a new optimized data set and then the above operations are repeated.

Expanding the corpus as depicted in Figure 3 can enhance the subsequent predictive capabilities. Specifically, by leveraging the features of professional data and utilizing reinforcement learning and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1



active learning mechanisms, effective solutions can be proposed for addressing the irregularities present in the repeatedly mentioned raw data. The core idea involves reordering internal communities and roads based on geographical information (communities and roads) as units and appropriately handling special characters. Then, an iterative process is carried out, and the determination of "iteration completion" is primarily based on the quantity or speed of generating new professional data. After each iteration, the need for further iterations is determined based on the situation of newly generated professional data. The iterative process is a crucial step in optimizing and expanding the corpus, significantly contributing to improving annotation quality.

## 4. Experiments and Applications

### *4.1. Automatic Annotation Main Module*

The original datasets and professional datasets can be obtained through existing tools or from existing databases. For example, they can be acquired through web scraping using a crawler. Each entry in the original dataset $UserInfSet$, represented by $UserInf$, signifies one piece of user information. For example, "韩@@ 138****7139 159****8976 商城路未来路交叉口东北角首座国际*号楼*单元**层**号 ( Han@@138****7139159***8976 At the northeast corner of the intersection of Shangcheng Road and Weilai Road, Building * Unit *, **th floor, Number **, Shouzu International.)" is one piece of user information. The professional dataset contains specialized data related to community information, such as the name of a community in a particular location, for example, “首座国际 (Shouzu International)”.

Due to the presence of duplicates or obvious errors in the acquired original dataset (UserInfSet), it is essential to conduct preprocessing to ensure the quality of the annotation work. Preprocessing steps like data filtering and cleaning are necessary. Additionally, since the professional dataset (PlotSet) may have inconsistent sources, such as data obtained from web scraping and various sources on the internet, it can lead to confusion and thus requires preprocessing as well.

The interface diagram of the automatic annotation system is shown in Figure 4.

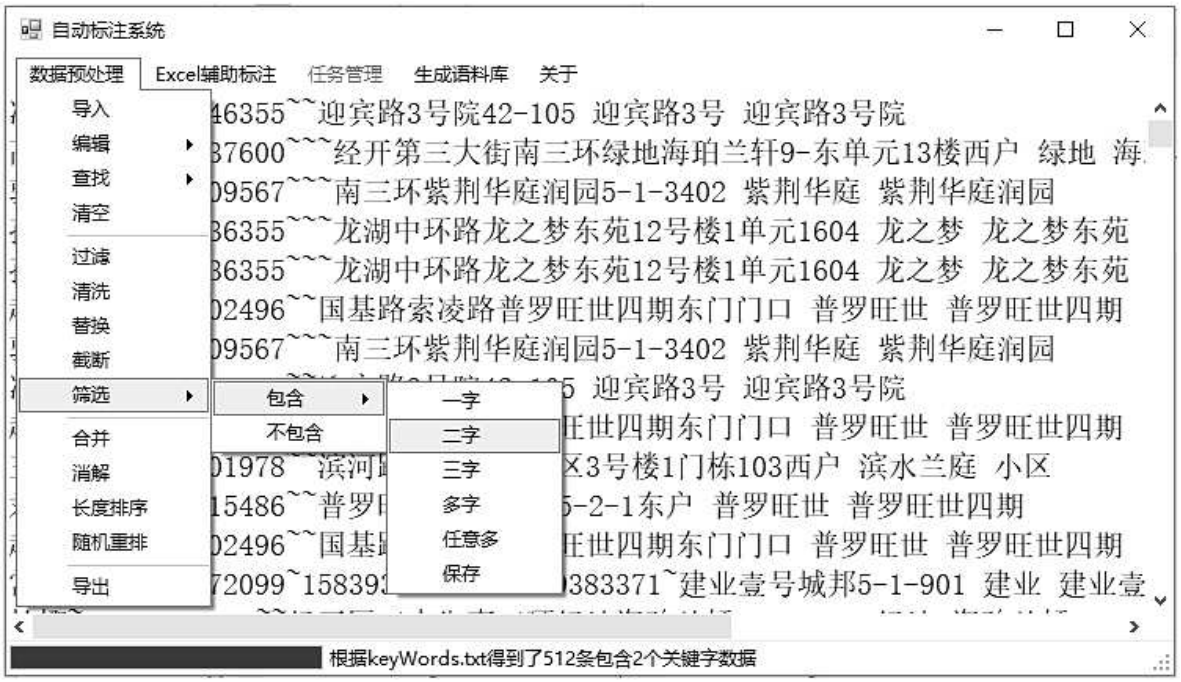


**Figure 4.** The interface diagram of the automatic annotation system.

The first-level menu that pops up in Figure 4 contains a full-featured preprocessing implementation. The second-level and third-level menus display the filtering functions containing

 doi:10.20944/preprints202309.1773.v1



professional data, allowing for the acquisition of higher-quality original data, thus implementing an active learning mechanism. Additionally, before assisting with manual annotation, the data is sorted by length to make the segmented data more suitable for manual annotation.

In the AGG layer, the geometric spaces in the merging function are divided into compatible, intersecting, adjacent, and separate. Geometric space describes the position relationship between two sub-segments located on the same line segment. Since this research pertains to community information, adjacency can be configured according to the actual situation. For example, situations where there are two characters in between but no specific word can also be considered adjacent. For instance, "绿地首府新区" where "绿地" and "新区" represent community information, while "首府" does not. In this case, the "adjacent" principle can be used to combine them, resulting in "绿地首府新区".

The default rules can be configured based on practical needs, such as sorting by record length, similarity, etc. The target dataset (SuperiorSet) is a dataset that contains target data, which are individual records. Random sorting can shuffle the order of the records, preparing them for subsequent steps.

The default identifier can be the number of cells, etc. which is primarily for grouping in subsequent manual adjustments. To facilitate manual optimization work, an Excel assistant design module has been developed (see Section 4.2), which places annotated data in cells on a per-character basis. By freezing the form, setting the font, interlacing the color, increasing the serial number of each word, and so on, the annotation work is further simplified, and the practice proves that the annotation work is well improved.

The automatically annotated data generated in this study is represented using Excel files. Afterward, the automatically annotated corpus is manually adjusted, and the quality of the manually adjusted annotated corpus is assessed. If it meets the standards, it is accepted through the evaluation; otherwise, manual adjustments continue. Next, corpus rules are configured, and based on the Excel files and the configured corpus rules, a corpus is synthesized. The synthesized corpus is then expanded, and the expanded corpus is input into the pre-built deep learning model (BiLSTM+CRF+AGG). The deep model with an aggregation layer (AGG) for training, results in a prediction model. The process then checks if the iteration is complete. If it is complete, the process ends; otherwise, the prediction model obtained from training is used to predict the original data not yet filtered into the target dataset. This provides expanded professional data. Finally, manual selection is conducted, and the selected data is integrated into the existing professional data.

### *4.2. Manual Annotation Assistance Module*

Practice has shown that automatic annotation can complete the majority of annotation work, and the results are quite satisfactory. However, there are still some complex cases that require manual correction [29], but the workload has been greatly reduced. At the same time, it is more valuable if the name of the cell whose *PlotCount* is 1 in step 3 of step 2 of the algorithm in 3.2 and the new cell obtained in the manual correction process is expanded to the professional data set. That is, better *LabelSet* and *PlotSet* can be obtained through continuous iteration.

To better facilitate the manual optimization work, an Excel-assisted design module has been developed. This module places the annotated data in cells on a per-character basis and simplifies the annotation process through functions such as freezing frames, adjusting fonts, alternating row colors, and adding sequence numbers for each character. The practice proves that the work of annotation is improved well, and the work of annotation is accurately evaluated through the evaluation module.

In the actual implementation process, the interface for importing lists and starting the generation of Excel annotation file sets is shown in Figure 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1




**Figure 5.** The interface diagram of manual assisted annotation.

Figure 6 is the interface for selecting the directory to generate Excel sets after importing the "preset identifiers." The "preset identifiers" are introduced for distinguishing tasks. The pre-processed, high-quality raw data is automatically segmented according to the preset identification, which also makes it easy to assign tasks. The mosaic in the figure is to protect user data privacy (the data shown in the figure is all real data). These data start with the name of the residential community and provide corresponding file names and line numbers, making it easier to locate information.


**Figure 6.** The illustrative interface for generating Excel-assisted files.

Figure 7 shows the generated Excel file. In Figure 7 each character has corresponding row and column numbers, and they are also color-coded to enhance readability and interest. To further reduce workload, numbers are used for annotation instead of letters. The Office's powerful features can also be utilized to freeze the annotation instructions and the first column.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1



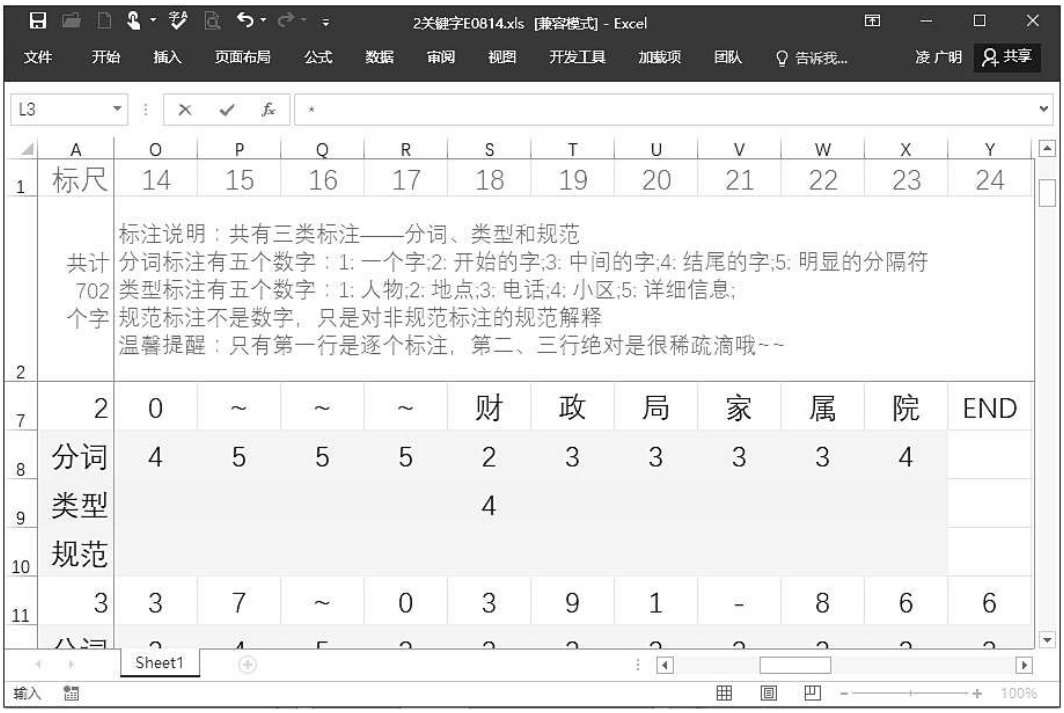


**Figure 7.** The Excel file for manual assistance annotation.

Furthermore, after further adjustments through manual annotation of N Excel files to obtain high-quality annotated data, it is necessary to assess the quality of the obtained annotations.

In the specific implementation process, the quality of annotations can be achieved by measuring the coverage of key annotations, such as the annotations for small areas. Due to the potential for errors during manual adjustments, a quality assessment is performed. The above-mentioned approach allows for precise assessment of the annotation work, further enhancing usability. As shown in Figure 8, the quality of the annotations is assessed by tallying problematic annotations and measuring their coverage.


**Figure 8.** A diagram of the assessment results.

 doi:10.20944/preprints202309.1773.v1



When annotating geographical information, international standards such as BIO, BIOS, and BIOES were employed. In this implementation, multiple annotation standards can be obtained through parameter design, and here the BIO standard was chosen. In the specific implementation, five categories of labels were defined, namely, to identify user names (PER), contact numbers (TEL), road information (LOC), neighborhood names (PLT), and neighborhood details (INF). Table 3 provides typical examples from five Chinese datasets that include geographical information.

**Table 3.** The examples of Chinese geographical names

| Corpus | Example | Chinese geographical name | Type |
|---|---|---|---|
| MSRA | 给孩子们以无私母爱的山西省大同市大同县倍加造镇解庄村村民任桂香 (Ren Guixiang, a villager from Jiezhuang Village, Beijiazao Town, Datong County, Datong City, Shanxi Province, who provides selfless motherly love to the children) | 山西省大同市大同县倍加造镇解庄村 (Jiezhuang Village, Beijiazao Town, Datong County, Datong City, Shanxi Province) | LOC |
| Resume | 李军，男，中国国籍，出生于1959年3月。自2008年12月起任中国工商银行股份有限公司非执行董事 (Li Jun, male, Chinese national, born in March 1959. He has been serving as a non-executive director at Industrial and Commercial Bank of China Limited since December 2008.) | 中国工商银行股份有限公司(Industrial and Commercial Bank of China Limited) | ORG |
| Weibo | 滨江森林公园之花朵篇单反没电，用拍的，没想到效果出奇好鼓掌我在 (In the flower section of Binjiang Forest Park, my DSLR camera ran out of battery, so I took photos with my smartphone. Surprisingly, the results turned out great! Applauding myself here.) | 滨江森林公园 (Binjiang Forest Park) | LOC.NAM |
| OntoNotes | 阿美达在7月份前往和落岛为遭到绑架、狭持的人质祷告 (Ahmada went to Heluo Island in July to pray for the hostages who were kidnapped and held hostage.) | 和落岛 (Heluo Island) | LOC |
| Address | 金华市婺城区宾虹西路1564号美保工具电商中心 (No. 1564 Binhong West Road, Wucheng District, Jinhua City, Zhejiang Province, China) | 整个句子(The entire sentence) | LOC |

### *4.3. Application of Research Results*

The three-dimensional scene geographic information perception system is based on the project "Key Project of National Key Research and Development Plan - Research on Common Basic Service Technology of Social Security Big Data". Through the cutting-edge deep learning technology in the field of artificial intelligence, the identification of the user's geographical location is studied and realized. The location is contained in unstructured text. The operating interface of the system is shown in Figure 9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2023 doi:10.20944/preprints202309.1773.v1



**Figure 9.** The application interface of research results.

In the dialogue box, you can input an event description, such as, "7 月22 日，在武汉龙王庙发生了爆炸。当日下午，浓烟滚滚。截止发稿时，大伙仍在继续，人员伤亡情况不详。近10 辆消防车已经前往补救 (On July 22nd, an explosion occurred at Longwang Temple in Wuhan. Thick smoke billowed into the air in the afternoon of the same day. As of the time of reporting, the situation was ongoing, and the extent of casualties was unknown. Nearly 10 fire trucks have been dispatched to provide assistance)." The application system developed in this research will automatically discover the geographical location information for "武汉龙王庙 (Longwang Temple in Wuhan)" from this unstructured text. This information is then submitted to the three-dimensional scene engine, which constructs the corresponding three-dimensional scene based on the provided data.

## 5. Conclusion

This research has implemented an automatic annotation system, integrating an active learning mechanism and an assisted manual annotation module to gradually optimize the annotation quality through an iterative process. Combining with the cutting-edge BiLSTM+CRF deep learning model and research on data characteristics, an improved model (BiLSTM+CRF+AGG) with an added aggregation layer, was proposed to achieve the optimal granularity for annotation geographical entities. To further enhance the model's generalization capabilities, data annotation was expanded, resulting in intelligent recognition of community names and geographic locations.

This research not only provides information services to society but also offers intelligent decision support for businesses. Future work should focus on further improving the sample selection and sequence generation methods and extend this model to broader scenarios through transfer learning.

**Author Contributions:** Conceptualization, Guangming Ling; Methodology, Aiping Xu; Software, Guangming Ling; Validation, Chao Wang; Resources, Guangming Ling and Chao Wang; Data curation, Guangming Ling; Writing—original draft preparation, Aiping Xu; Writing—review and editing, Guangming Ling; funding acquisition, Chao Wang. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Key R&D Program of China (No. 2018YFB2100603), the Key R&D Program of Hubei Province (No. 2022BAA048), the National Natural Science Foundation of China program (No. 41890822) and the Open Fund of National Engineering Research Centre for Geographic Information System, China University of Geosciences, Wuhan 430074, China (No. 2022KFJJ07). The numerical calculations in this paper have been done on the super-computing system in the Supercomputing Centre of Wuhan University.

**Data Availability Statement:** Not applicable.

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

## References


1. Li, H.; Lu, W.; Xie, P.; Li, L. Neural Chinese Address Parsing. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp. 3421–3431.
2. Yassine, M.; Beauchemin, D.; Laviolette, F.; Lamontagne, L. Multinational Address Parsing: A Zero-Shot Evaluation. *6*, 40–50. doi:10.57675/IMIST.PRSM/ijist-v6i3.187.
3. Du, L.; Xiao, G. Automatic Entity Recognition of Geographic Information Service Document. *Natural Science Journal of Hainan University* **2021**, *39*, 331–338. doi:10.15886/j.cnki.hdxbzkb.2021.0042.
4. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5847–5858. doi:10.18653/v1/2021.acl-long.454.
5. He, Y.; Luo, C.; Hu, B. Geographic Entity Recognition Method Based on CRF Model and Rules Combination. *Computer Applications & Software* **2015**, *32*, 179–185+202. doi:10.3969/j.issn.1000-386x.2015.01.046.
6. Graves, A.; Mohamed, A.r.; Hinton, G. Speech Recognition With Deep Recurrent Neural Networks. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; IEEE, , 2013; pp. 6645–6649.
7. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
8. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, D. Neural architectures for named entity recognition. Proceedings of NAACL-HLT'16; IEEE Press, , 2016; pp. 260–270.
9. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* **2015**.
10. Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; Di, H. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. Natural Language Understanding and Intelligent Applications; Springer International Publishing: Cham, 2016; pp. 239–250.
11. Shen, Y.; Yun, H.; Lipton, Z. Deep active learning for named entity recognition. Proceedings of the 2nd Workshop on Representation Learning for NLP; Association for Computational Linguistics, , 2017; pp. 252–256.
12. Yang, Y.; Zhang, W. Image Auto-annotation Based on Deep Learning. *Journal of Data Acquisition and Processing* **2015**, *30*, 88–98. doi:10.16337/j.1004-9037.2015.01.008.
13. Xu, Y.; Zhang, H. Summary of automatic image annotation method. *Journal of Modern Information* **2016**, *36*, 144–150. doi:10.3969/j.issn.1008-0821.2016.03.024.
14. Cheng, B. The study of automatic labeling technology based on convolution neural network. *Electronics World* **2019**, pp. 124–126.
15. Huang, D.; Xu, Q.; He, Q.; Du, M. Multi-features fusion for image auto-annotation based on DBN model. *Computer Engineering and Applications* **2018**, *54*, 224–228. doi:10.3778/j.issn.1002-8331.1607-0297.
16. Huang, D.; Xu, Q.; He, Q.; Du, Y. Multi-features fusion for image auto-annotation based on DBN model. *Computer Engineering and Applications* **2018**, *54*, 224–228. doi:10.3778/j.issn.1002-8331.1607-0297.
17. Qiu, P.; Zhang, H.; Yu, L.; Lu, F. Automatic Event Labeling for Traffic Information Extraction from Microblogs. *Journal of Chinese Information Processing* **2017**, *31*, 107–114.
18. Chou, P.; Lu, F.; Zhang, H.; Yu, L. Automatic identification method of micro-blog messages containing geographical events. *Journal of Geo-information Science* **2016**, *18*, 886–893. doi:10.3724/SP.J.1047.2016.00886.
19. Xu, F.; Ye, W.; Song, Y. Part-of-speech automated annotation of food safety events based on BiLSTM-CRF. *Journal of the China Society for Scientific and Technical Information* **2018**, *37*, 1204–1211. doi:10.3772/j.issn.1000-0135.2018.12.004.
20. Wang, J.; Lu, F. Constructing the corpus of geographical entity relations based on automatic annotation. *Journal of Geo-Information Science* **2018**, *20*, 871–879. doi:10.12082/dqxxkx.2018.180032.
21. Zhu, Z.; Li, S.; Dai, M.; Zhou, G. Opinion target extraction with active-learning and automatic annotation. *Journal of Shandong University* **2015**, pp. 38–44. doi:10.6040/j.issn.1671-9352.3.2014.106.
22. Schulz, C.; Meyer, C.; Kiesewetter, J.; Sailer, M.; Bauer, E. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics, , 2019; pp. 2761–2772.

23. Wu, K.; Zhang, X.; Ye, P.; Hua, A.; Zhang, H. A Chinese Address Resolution Method Based on BERT-BiLSTM-CRF. *Geography and Geoinformation Science* **2021**, *37*, 10–15.
24. LING, G.m.; XU, A.p.; WANG, W. Research of Address Information Automatic Annotation Based on Deep Learning. *ACTA ELECTONICA SINICA* **2020**, *48*, 2081–2091. doi:10.3969/j.issn.0372-2112.2020.11.001.
25. Ling, G.; Mu, X.; Wang, C.; Xu, A. Enhancing Chinese Address Parsing in Low-Resource Scenarios through In-Context Learning. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 296. doi:10.3390/ijgi12070296.
26. Ling, G.; Xu, A.; Wang, C.; Wu, J. REBDT: A Regular Expression Boundary-Based Decision Tree Model for Chinese Logistics Address Segmentation. *Appl. Intell.* **2023**, *53*, 6856–6872. doi:10.1007/s10489-022-03511-6.
27. University, W. An automatic labeling method and device based on in-depth learning: CN201811434810.2, 12, 2020.
28. Hu, Y.; Zheng, X.; Zong, P. An Active Transfer Learning Method Combining Uncertainty with Diversity for Chinese Address Resolution. Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition. Association for Computing Machinery, ICCPR '22, pp. 643–650. doi:10.1145/3581807.3581902.
29. McDonald, Y.J.; Schwind, M.; Goldberg, D.W.; Lampley, A.; Wheeler, C.M. An analysis of the process and results of manual geocode correction. *Geospatial health* **2017**, *12*, 526.