Research on Named Entity Recognition Methods for Urban Underground Space Disasters Based on Text Information Extraction

Zhaowen Li[1], Xuedong Zhang[1,2*]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China - 2108570021075@stu.bucea.edu.cn, zhangxuedong@bucea.edu.cn
[2] Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing 100038, China

KEY WORDS: Urban Underground Space Disasters, Named Entity Recognition, ALBERT, BiLSTM, CRF.

ABSTRACT:

Urban underground space is a complex spatial scenario that is highly susceptible to disasters. To achieve entity recognition in textual information related to urban underground space disasters, this study proposes the ALBERT-BiLSTM-CRF model. The urban underground space disaster text data is firstly encoded using the ALBERT model, which captures the deep semantic information of words in the context. The encoded data is then fed into a BiLSTM network to obtain hidden state vectors for each word, enhancing the feature representation of words. Finally, these vectors are input into a CRF layer to obtain the optimal label sequence and complete named entity recognition. The proposed model achieves an accuracy of 95.41%, a recall of 94.08%, and an F1 score of 94.74%. Comparative experiments with the BiLSTM-CRF, BERT-CRF, and BERT-BiLSTM-CRF models are conducted on the Boson dataset and our experimental dataset, demonstrating the superior performance of the ALBERT-BiLSTM-CRF model.

1. INTRODUCTION

With the continuous exploration of modern cities into underground spaces, the importance of urban underground space in the urban economy, transportation, and other aspects has become increasingly prominent. Urban underground space includes transportation facilities, municipal engineering facilities, commercial facilities, disaster prevention facilities, and more. Urban underground space is also a highly susceptible scenario for disasters. Based on statistical results of urban underground space disaster types over the past few decades, it has been found that flooding, earthquakes, fires, explosions, deformations, cracks, and leaks are typical and frequent disasters in urban underground spaces. Most of the urban underground space disaster events are stored in text form, and extracting useful information from a large amount of redundant Chinese textual data requires significant human and material resources. Constructing a knowledge graph as the mainstream solution, analyzing a large amount of redundant text data using the analysis methods provided by knowledge graphs, constructing knowledge bases, and returning the required information to users.

Named Entity Recognition (NER) is the most fundamental and critical step in constructing a knowledge graph. Its purpose is to identify named entities with specific meanings from text, such as personal names, place names, organization names, time, dates, and more. The task of NER is to automatically extract and analyze a large amount of text data to discover important information and relationships, which is conducive to the automatic processing and mining of information, and then identify named entities related to urban underground space disasters from the text. These named entities include the following types:
(1) Underground spaces: including underground tunnels, subways, underground shopping malls, underground parking lots, etc.
(2) Disaster types: including fires, gas leaks, floods, earthquakes, etc.
(3) Loss situations: including casualties, property damage, etc.
(4) Rescue measures: including rescue teams, rescue equipment, rescue time, etc.

NER methods can be categorized into rule-based and dictionary-based approaches, supervised learning methods, and deep learning methods. In recent years, there has been a growing interest in deep learning-based NER methods. Compared to traditional machine learning, deep learning facilitates the automatic discovery of hidden features and feature extraction, thereby improving generalization capabilities (Li, Sun, Han, & Li, 2022). Achieving good performance in NER, researchers have gradually improved neural network structures to meet the requirements of NER tasks. For instance, Lample et al. (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016) proposed the BiLSTM-CRF model, which incorporates a CRF layer to optimize the output label sequence and achieved a high F1 score in the corresponding corpus. Ma et al. (Ma & Hovy, 2016) introduced the BiLSTM-CNNs-CRF model by combining the bidirectional LSTM and CNN structures with a CRF layer, resulting in improved F1 scores. Luo et al. (Luo et al., 2018) proposed the Att-BiLSTM-CRF model, which is capable of document-level NER and achieved an F1 score of 91.14% on the corresponding dataset. These models primarily focus on word or character feature extraction and do not consider dynamic word semantics or word ambiguity in the contextual context.

To address these limitations, Devlin et al. (Devlin, Chang, Lee, & Toutanova, 2018) introduced the BERT model, a bidirectional language model that can better understand vocabulary and grammar structures in the context, thereby handling complex natural language tasks more effectively. Xie et al. (Teng, Jun-An, & Hui, 2020) proposed a BERT-BiLSTM-CRF model and conducted experiments on two corpora, achieving F1 scores of 94.65% and 95.67%, respectively. Baidu developed the ERNIE model (Sun et al., 2019), an improvement upon the BERT model, which includes additional pre-training tasks and introduces adaptive attention mechanisms, enhancing the model's expressive

https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-547-2023 | © Author(s) 2023. CC BY 4.0 License.

power. Microsoft proposed the MT-DNN model (Liu, He, Chen, & Gao, 2019), which employs a multi-task training approach and demonstrates greater stability and generalization capabilities than BERT. Google introduced the ALBERT model (Lan et al., 2019), an improved version of the BERT model, with techniques such as parameter sharing and cross-layer parameter loss, reducing the number of model parameters while improving training efficiency and generalization capabilities.

Currently, deep learning-based NER has been extensively applied in various domains. For instance, Du et al. (Jin-hu, Hao, & Song, 2022) conducted NER research on Chinese electronic medical records to support medical intelligence. Li et al. (Dongsheng, Yulai, Jianhua, & Dewang, 2023) focused on NER in library service information and utilized the BERT-BiLSTM-CRF model, achieving an F1 score of 98.75% to advance intelligent services in university libraries. While NER has been widely applied in different fields, there is a lack of research specifically targeting NER for urban underground space disaster knowledge.

This study focused on Chinese textual data related to urban underground space disaster events and proposed the construction of an ALBERT + BiLSTM + CRF model to achieve NER of typical urban underground space disaster knowledge. The model eliminated the labour-intensive manual feature extraction methods. The ALBERT model provided embedding vectors for the input sequence, representing the feature representation of each word or character in the Chinese textual data. The BiLSTM model captured contextual information from both preceding and succeeding sequences, enhancing the contextual relevance of each word or character. The CRF model decoded the encoded sequence to obtain the optimal label sequence.

2. RELATED WORK

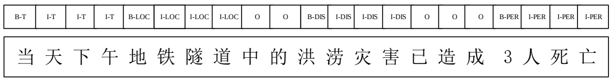
2.1 Data Preprocessing

The data for this study primarily comes from four sources. Firstly, Chinese academic websites such as CNKI were utilized to gather relevant literature. Secondly, open datasets from general knowledge sources, including encyclopedia websites and DBpedia, were collected. Thirdly, news portals related to urban underground space disasters were mined for data. Lastly, data accumulated by research institutions and relevant books were included. Data collection involved manual searching and web scraping techniques. The collected text data underwent noise reduction preprocessing using Python scripts. Approximately 1 million Chinese characters were retained for analysis. As shown in Figure 1, the HIT LTP (Harbin Institute of Technology Lexical Analysis Tool) was used for word segmentation, enabling the analysis of urban underground space disaster text data. Through manual annotation, entities related to urban underground space, disaster types, losses, and rescue measures were identified and added to the LTP word segmentation tool as a dictionary.

Figure 1. Original text data of urban underground space hazards and text data after word separation

2.2 Part-of-Speech Tagging

Part-of-speech (POS) tagging involves assigning a grammatical category to each word in a text, such as noun, verb, adjective, etc. In the task of NER, POS tagging serves as an auxiliary tool for identifying entity types and contextual relationships. In this study, the BIO (Beginning, Inside, Outside) method was adopted for POS tagging. The "B" denotes the beginning of an entity, and the "I" represents all labels following the starting node. The LTP tool was used for POS tagging, and to improve accuracy, entity types were abstracted into five categories: (1) Location (LOC), including underground passages, subways, underground shopping malls, underground parking lots, and other location information; (2) Person (PER), including personal names and information about casualties; (3) Time (T), representing all time-related information; (4) Organization (ORG), encompassing rescue organizations, social organizations, and other institutions; (5) Disaster (DIS), comprising information about types of underground space disasters and the resulting losses. Figure 2 illustrates the sequence labelling format.



B-T	I-T	I-T	I-T	B-LOC	I-LOC	I-LOC	I-LOC	O	O	B-DIS	I-DIS	I-DIS	I-DIS	O	O	O	B-PER	I-PER	I-PER	I-PER
当	天	下	午	地	铁	隧	道	中	的	洪	涝	灾	害	已	造	成	3	人	死	亡

Figure 2. Sequence annotation

3. RESEARCH METHODOLOGY

3.1 ALBERT Model

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model based on the Transformer architecture (Vaswani et al., 2017). It possesses rich language representation capabilities and can be applied to various natural language processing tasks. Figure 3 illustrates the architecture of the BERT model. The pre-training process of BERT is conducted in an unsupervised manner, where the model learns contextual representations of sentences by predicting missing words from a large amount of unlabelled text. The main contribution of the BERT model is the incorporation of bidirectional contextual information into the pre-training process, enabling a better understanding of semantics and grammar in natural language(Rogers, Kovaleva, & Rumshisky, 2020).

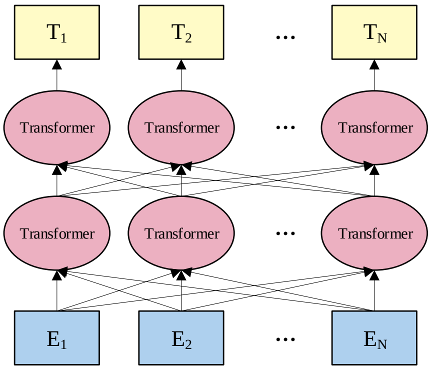


Figure 3. BERT model architecture

A Lite BERT (ALBERT) is a lightweight model based on the BERT model. Similar to BERT, the ALBERT model employs the Transformer architecture for language modelling. However, ALBERT introduces several techniques to improve the efficiency and performance of the BERT model. Firstly, ALBERT utilizes the technique of word-piece embeddings, which splits the vocabulary embedding matrix of BERT into multiple smaller matrices. This allows multiple words to share the same small matrix, reducing the number of model parameters and improving efficiency. Secondly, ALBERT incorporates cross-layer parameter sharing, breaking down the parameters of each layer in BERT into smaller blocks and sharing these blocks across layers, further reducing the model's parameter count. Additionally, ALBERT incorporates other techniques such as sentence order prediction and downsampling to enhance both efficiency and performance (Zhang, Li, & Du, 2020). In comparison to the BERT model, the main differences of the ALBERT model lie in its parameter count and training efficiency. ALBERT has only half or even fewer parameters than BERT, while achieving higher training efficiency. In most natural language processing tasks, the ALBERT model outperforms the BERT model in terms of various performance metrics and is more efficient in terms of parameter count and training time.

3.2 BiLSTM Model

The Bidirectional Long Short-Term Memory (BiLSTM) model is a neural network model used for processing sequential data and is a variant of the Long Short-Term Memory (LSTM) model. The LSTM model is a special type of recurrent neural network (RNN) that can handle long-term dependencies in sequential data, overcoming the vanishing gradient problem commonly encountered in traditional RNN models. As RNN networks transmit all information to the next layer during runtime, it is challenging to effectively capture long-term dependencies. Figure 4 illustrates the architecture of the LSTM model, which incorporates forget gates, input gates, and output gates to address this issue.

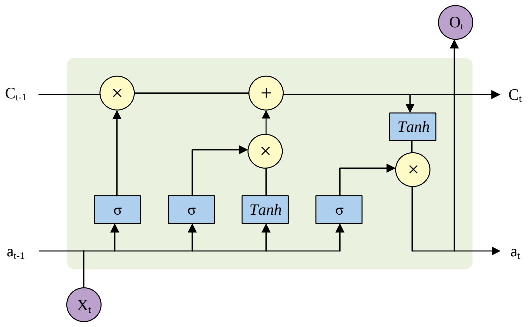


Figure 4. LSTM model architecture

The BiLSTM model extends the LSTM model by adding an additional LSTM layer that propagates information in the reverse direction, as shown in Figure 5. This allows the model to simultaneously consider both forward and backward information in the input sequence, resulting in a better understanding of the contextual information within the sequence. The word vectors generated by the ALBERT model are processed through a BiLSTM layer to obtain contextual feature representations for each word. The role of the BiLSTM model is to capture the contextual information of words within the entire sentence, enabling a better understanding of word meanings and semantics (Luo et al., 2018).

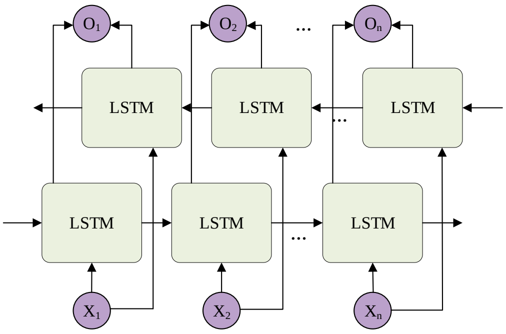


Figure 5. BiLSTM architecture

3.3 CRF Model

Conditional Random Field (CRF) is a probabilistic model used for sequence labelling. Figure 6 illustrates the architecture of the CRF model, which is a special case of a Markov Random Field. In CRF, it is assumed that there are only two variables, X and Y, and X is known while Y is the output given X. Here, $X=(X_1,X_2\dots X_n)$, $Y=(Y_1,Y_2\dots Y_n)$,Given an input sequence X, the conditional probability of predicting an output sequence Y is calculated as follows:

$$\mathrm{P(Y|X)} = \frac{1}{Z(X)}\exp\left(\sum_{l,k}\lambda_k t_k(Y_{i-1},Y_i,x,i)+\sum_{l,i}\mu_l s_l(Y_i,x,i)\right) \tag{1}$$

$$\mathrm{Z(X)} = \sum_{Y}\exp\left(\sum_{l,k}\lambda_k t_k(Y_{i-1},Y_i,x,i)+\sum_{l,i}\mu_l s_l(Y_i,x,i)\right) \tag{2}$$

In the equation, t_k, s_l represent feature functions. $t_k$refers to the transition features defined on edges, which depend on the current and previous positions. $s_l$represents the state features defined on nodes, which depend on the current position. λ_k, μ_l are the corresponding weight values.

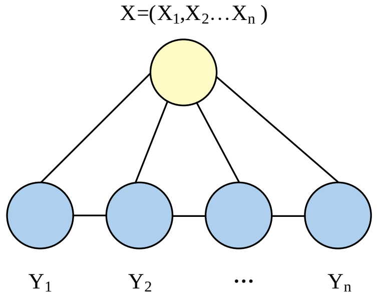


Figure 6. CRF model architecture

In the task of NER, the BiLSTM model models the contextual information of each word and outputs a score matrix containing various possible label sequences. These label sequences represent the possible labels for each word (e.g., person name, location name, organization name, etc.), and the score of each label sequence indicates the likelihood of its occurrence in the current sentence. The CRF model utilizes these scores along with the transition probabilities between labels in each label sequence to calculate the most probable label sequence (Jin et al., 2019). This enables the conversion of the probability matrix output by the BiLSTM model into the final annotation result, thereby improving the accuracy and robustness of the model.

3.4 ALBERT-BiLSTM-CRF Model

As shown in Figure 7, the model consists of an ALBERT layer, a BiLSTM layer, and a CRF layer from bottom to top. The ALBERT layer serves as a feature extractor, transforming the input urban underground space disaster text data into a high-dimensional vector representation to extract semantic information, which helps the model better understand the contextual relationships in the text. Then, through the bidirectional LSTM (BiLSTM) layer, the contextual information of the text data is modelled to capture long-term dependencies in the text. Finally, the conditional random field (CRF) layer globally optimizes the output label sequence of the model to make the predicted label sequence more accurate.

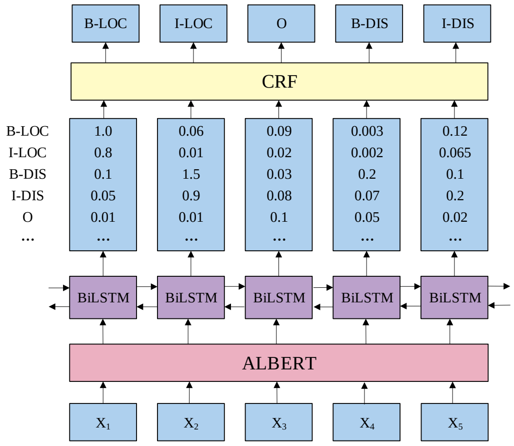


Figure 7. ALBERT-BiLSTM-CRF model architecture

4. EXPERIMENT AND RESULTS

4.1 Model Evaluation Metrics

The ALBERT-BiLSTM-CRF model was evaluated using precision, recall, and F1 score as the evaluation metrics during training. Precision is the ratio of correctly predicted positive samples to the total predicted positive samples, and it serves as a measure of the model's accuracy. Recall is the ratio of correctly predicted positive samples to the total actual positive samples, and it measures the model's ability to capture all positive instances (Table 1). The F1 score is the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance. The formulas are as follows:

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F}_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

Model	Actual Positive Example	Actual Negative Example
Predicted Positive	TP	FP
Predicted Negative Cases	FN	TN

Table 1. Classification results table

4.2 Experimental Setup and Parameter Settings

The experiments were conducted using the Tensorflow 1.14 framework and Python 3.7 environment. The GPU used was Nvidia RTX 3060 with 12GB of memory. The pre-trained model employed was ALBERT_base_zh, which has a parameter size of 12MB and consists of 12 layers, resulting in a total size of 40MB. For the NER task, the model was trained with the following parameters after preprocessing the Chinese text data and obtaining 18,109 segmented sentences. The configuration of other model parameters is presented in Table 2.

Parameter	Parameter Value
max_length	202
lr	0.00001
embedding_dim	128
rnn_units	128
dropout	0.5
vocab_size	21128
batch_size	16

Table 2. Parameter configuration

4.3 Experimental Results

Due to the large workload of manual annotation, we conducted experiments using a combined dataset consisting of manually labelled data and a similar dataset from the People's Daily. The entity annotation dataset was merged in a 1:1 ratio, resulting in a final dataset size of 3.1MB. The dataset was then divided into training, validation, and testing sets in an 8:1:1 ratio for experimentation purposes.

Entity Type	Precision	Recall	F_1	Dataset size
LOC	95.67%	94.59%	95.13%	3582
ORG	92.33%	93.09%	92.71%	2203
PER	98.59%	97.69%	98.14%	1842
DIS	96.67%	95.23%	95.94%	1112
T	83.03%	86.34%	84.65%	673

Table 3. Named entity identification results

According to Table 3, the model achieved high F1 scores in the location entity, population entity, and disaster entity categories, reaching 98.14%. This is mainly due to the fixed and well-defined nature of entity nouns in the location, population, and disaster types. However, the model performed poorly in the organization entity and time entity categories, which can be attributed to the following reasons: (1) The types and names of organization entities are frequently updated, and the manually labelled dataset is limited. (2) Time entities were annotated considering relative terms such as "on that day" or "subsequently", which lack clear boundaries for time entity nouns. (3) The time entity category had fewer annotations compared to the other four categories, which to some extent affected the entity recognition results. To address these issues, it is suggested to expand the dataset through manual augmentation or data augmentation techniques.

4.4 Model Comparison Experiment and Analysis

To verify the advantages of the ALBERT-BiLSTM-CRF model, we conducted comparative experiments with three commonly used models in the current NER task: BiLSTM-CRF, BERT-CRF, and BERT-BiLSTM-CRF. Considering the complexity of the urban underground space disaster NER task, we selected the open-source Boson dataset for the comparative experiment. The Boson dataset includes six entity types: company_name, Location, org_name, person_name, product_name, and time, which are similar to the types of underground space disaster entities. The four models were configured with the same parameters, including a batch size of 16, 15 epochs, and a learning rate of 0.00001. The experimental results are shown in Table 4, clearly indicating that the ALBERT-BiLSTM-CRF model outperformed the other models, with BERT-BiLSTM-CRF performing second-best and BiLSTM-CRF having the lowest performance.

Model	Precision	Recall	F_1
BiLSTM-CRF	77.94%	82.23%	80.03%
BERT-CRF	89.37%	90.42%	89.89%
BERT-BiLSTM-CRF	89.93%	91.70%	90.81%
ALBERT-BiLSTM-CRF	92.71%	91.83%	92.27%

Table 4. Boson dataset experiment results

The urban underground space dataset was inputted into the four models, with the same parameters, and the results are shown in Table 5. All four models showed improved performance compared to the Boson dataset. This improvement can be attributed to the larger number of entity types in the Boson dataset, including entities such as company names and product names that are more challenging to recognize. Among the four models, the ALBERT-BiLSTM-CRF model performed the best in the urban underground space disaster dataset recognition task, exhibiting a significant improvement over the BERT-BiLSTM-CRF model. This can be attributed to the improvements made to the BERT model through a series of techniques in the ALBERT model, resulting in higher efficiency and better performance. Thus, it can be concluded that utilizing the ALBERT-BiLSTM-CRF model for urban underground space disaster NER tasks is the optimal approach.

Model	Precision	Recall	F_1
BiLSTM-CRF	90.06%	89.13%	89.59%
BERT-CRF	92.43%	91.31%	91.87%
BERT-BiLSTM-CRF	92.34%	92.18%	92.26%
ALBERT-BiLSTM-CRF	95.41%	94.08%	94.74%

Table 5. Experimental results of urban underground space disaster dataset

5. CONCLUSION

The proposed ALBERT-BiLSTM-CRF model is a deep learning model for NER. It consists of three components: the ALBERT pre-trained language model, bidirectional long short-term memory (BiLSTM), and conditional random field (CRF). The ALBERT-BiLSTM-CRF model takes the text sequence as the input in the NER task, extracts feature representations using the pre-trained ALBERT model, performs sequence modelling through the bidirectional LSTM, and finally labels the output using the CRF model. The model demonstrated good performance on the open-source NER datasets and achieved satisfactory results on the urban underground space disaster dataset. The accuracy of the model reached 95.41%, the recall rate reached 94.08%, and the F1 value reached 94.74 %. These experimental results indicated that the model is superior to some commonly used NER methods, and provides valuable insights for NER tasks in the field of urban underground space. However, it is worth noting that the experiment only collected a limited amount of event and literature data related to urban underground space disasters, which introduces limitations in the diversity of the NER task data. Future work will focus on expanding the types of urban underground space disaster data to enhance the applicability of the ALBERT-BiLSTM-CRF model and support the construction of a knowledge graph for urban underground space disasters.

ACKNOWLEDGMENTS

This study was supported by grants from the following projects: Beijing Key Laboratory of Urban Spatial Information Engineering (NO. 20230101), Beijing Higher Education Undergraduate Teaching Reform and Innovation Project (NO. 202110016004) and Beijing Education Science Planning Project (NO. CDDB21191).

REFERENCES

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dongsheng, L., Yulai, B., Jianhua, L., & Dewang, C. (2023). Named Entity Recognition Method of the University Library Wechat Information Service Based on BE Ｒ T. *Journal of Modern Information, 43*(04), 64-76.

Jin-hu, D., Hao, Y., & Song, F. (2022). Research and Development of Named Entity Recognition in Chinese Electronic Medical Record. *ACTA ELECTRONICA SINICA, 50*(12), 3030-3053.

Jin, Y., Xie, J., Guo, W., Luo, C., Wu, D., & Wang, R. (2019). LSTM-CRF neural network with gated self attention for Chinese NER. *IEEE Access, 7*, 136694-136703.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Li, J., Sun, A. X., Han, J. L., & Li, C. L. (2022). A Survey on Deep Learning for Named Entity Recognition. *Ieee Transactions on Knowledge and Data Engineering, 34*(1), 50-70.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics, 34*(8), 1381-1388.

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics, 8*, 842-866.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Wu, H. (2019). Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Teng, X., Jun-An, Y., & Hui, L. (2020). Chinese Entity Recognition Based on BERT-BiLSTM-CRF Model. *Computer Systems & Applications, 29*(07), 48-55.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Zhang, X., Li, C., & Du, H. (2020). *Named Entity Recognition for Terahertz Domain Knowledge Graph based on Albert-BiLSTM-CRF*. Paper presented at the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-547-2023