Research on the Construction and Application of Knowledge Graph in the Field of Open Data Policy

Gao Lu[1], Xingli Liu[2(✉)], Jianghong Ou[3], and Dahua Fan[3]

[1] Harbin Institute of Information Technology, No.9, University Town, Binxi Technological Development Zone, Harbin 150431, Heilongjiang, China
[2] School of Computer Science and Technology, Science and Technology, University of Heilongjiang, Harbin 150027, Heilongjiang, China
liuxingli@usth.edu.cn
[3] Starway Communication, No. 31, Kefeng Road, Guangzhou Science City, 510663 Guangzhou, China

Abstract. In this paper, data open policy documents, laws and regulations are used as corpus sources, and the Bi-LSTM + CRF deep learning algorithm is selected to complete the training of the named entity recognition model constructed by the knowledge graph, and realize a collaborative relationship, data openness and data security concepts as the ontology. The knowledge map in the field of data openness policy is used to construct the model to complete the automatic identification and analysis of the collaborative situation of data openness policy text. The final simulation verification shows that the Bi-LSTM + CRF named entity recognition algorithm is more accurate than the CRF + + machine learning model training accuracy P value, recall rate R value and the harmonic average F value have been significantly improved, and the "Outline for Promoting Big Data Development", a typical data development policy text coordination situation analysis, has been objectively completed from the perspective of data openness and data security.

Keywords: Policy Text Analysis · Named Entity Recognition · Bi-LSTM + CRF · Knowledge Graph · Data Open Policy

1 Introduction

Today, countries around the world have continuously improved their awareness of the value of data. Developed countries in Europe and the United States have formulated a series of policies related to big data. The United States has issued the Federal Big Data Research and Development Strategic Plan [1], and the United Kingdom has issued the Industrial Strategy: Artificial Intelligence. "Domain Action" [2]. Subsequently, the

This research is financially supported by the National Social Science Foundation of China (GrantNo. 20ATQ004).

X. Jiang (Ed.): MLICOM 2022, LNICST 481, pp. 137–147, 2023.
https://doi.org/10.1007/978-3-031-30237-4_13

analysis of open government data policy has become a research hotspot. For example, Ma Haiqun [3] proposed a scientific method system based on policy, and conducted research on open government data policy collaboration from multi-dimensional perspectives such as policy elements, policy processes, and policy categories. In recent years, with the increase of policy texts and the rapid development of artificial intelligence technology, the application of automatic semantic extraction methods for policy texts in specific application scenarios has gradually become a hot and difficult point. The challenge lies in how to use information science, computer science, and literature information. A comprehensive research method is used to complete the process of policy text analysis of automatic human-computer interaction, so as to interpret the semantics of a core issue in the policy text through computer intelligence [4]. This non-intrusive and imprecise research method is the process of conducting natural language processing of policy texts in an objective and neutral position, and realizing the cognitive and intelligent recognition of policy texts by computers [5]. This paper takes the "data openness" policy text document data as the research object, applies the knowledge graph structured knowledge representation, adopts the Bi-LSTM + CRF named entity recognition deep learning algorithm to perform fine-grained custom entity object extraction training, and uses "Take the Outline for Promoting the Development of Big Data as an example to complete the application analysis of collaborative situation, and explore the feasibility of an intelligent automatic extraction method of data opening policy text".

2 Related Research

Reviewing the research results and literature of different policy text analysis, it is found that by completing the mining of policy text information and the analysis of external attribute characteristics through quantitative statistical tools, objective and verifiable research conclusions can be obtained [6]. This kind of policy literature measurement is a The current main policy analysis method is the organic integration of bibliometric analysis and content analysis [7, 8]. Aiming at the results of quantitative method research on data open policy, Chu Dejiang, with the help of Nvivo12 qualitative data analysis software, followed the steps of analysis framework construction, policy text coding, frequency statistical analysis, etc., from the two dimensions of policy instrumentality and synergy. 33 A quantitative text analysis was carried out on the policies closely related to rural green development [9]; Yang Zheng [10] and others applied policy bibliometric methods to analyze the significance of China's data openness and utilization policy system for promoting data governance and mining data value; Jiang Xin [11] Evaluated and analyzed the open and shared policies of scientific data issued by foreign funding agencies through qualitative text analysis and put forward suggestions; for the results of policy collaborative research, Hong Weida [12] et al. For government policy, policy quantification standards were designed from three dimensions of policy strength, policy objectives and policy tools, and the collected policy texts were quantified by grades, and a measurement model of policy effectiveness, policy objective synergy degree, and policy tool synergy degree was constructed. From the time dimension to analyze the policy synergy degree of China's open government data; Zhang Tao [13] analyzed the theme synergy degree of 446 policy texts based on the policy text calculation method, and

used the LDA theme clustering method to obtain the policy text theme synergy degree value; Mao Zijun [14] et al. took 12 provinces and cities data opening related policies as research samples, analyzed from two dimensions of vertical policy coordination and horizontal policy coordination, and revealed the policy coordination between cities; Chen Xuelin [15] et al. From the perspective of scientific knowledge map, Based on co-word analysis, the research on hotspots of entrepreneurship and innovation policies in China combined with the results of co-word clustering analysis, using multi-dimensional scale analysis method, to draw a knowledge map of entrepreneurship and innovation policies, and explore the closeness of keywords in the hotspot areas of entrepreneurship and innovation policies and the hotspots. The relationship between fields, and finally found the internal structural relationship of China's mass entrepreneurship and innovation policy.

Based on the above analysis, it can be seen that the analysis methods at the textual level such as knowledge graph construction and related named entity recognition are feasible. Therefore, this paper takes the data opening policy text as the research object, determines the methods of ontology construction and entity extraction named entity recognition in the construction of knowledge graph, analyzes the synergy of policy text from the semantic analysis level, and completes the analysis of domain knowledge graph in data opening policy text.

3 Research Methods and Results

3.1 Overall Framework

According to the general policy text analysis process, according to the policy text acquisition (Acquire Documents), policy text processing (Process), policy text analysis (Analysis), construct the data opening policy text synergy analysis framework, as shown in Fig. 1.

① Acquire Documents. Policy text acquisition is the premise and foundation of policy text calculation. The policy text acquisition of this model needs to obtain data policy corpus through CNKI, policy documents, etc., and conduct corpus screening.

② Policy text processing (Process). The policy text processing process mainly includes the following: corpus collection, knowledge representation, policy text segmentation, part-of-speech tagging, domain dictionary construction, and using this dictionary to expand the Jieba vocabulary of the Chinese word segmentation tool, and segment the corpus to provide data input for model training. After superimposing and quality noise processing for the number of policy corpora, the qualified corpus is converted into data format, including BIO data format, Word2vec word vector processing, and divided into training set, development set and test set according to 7:2:1 and imported into the model, set the number of model iterations, and complete the Bi-LSTM + CRF named entity recognition model training process.

③ Policy text analysis (Analysis). First of all, it is necessary to encapsulate the named entity recognition algorithm based on knowledge graph through the visualization platform, select the data opening policy text that needs to be automatically parsed, and conduct calculation and supply policy managers for application analysis.

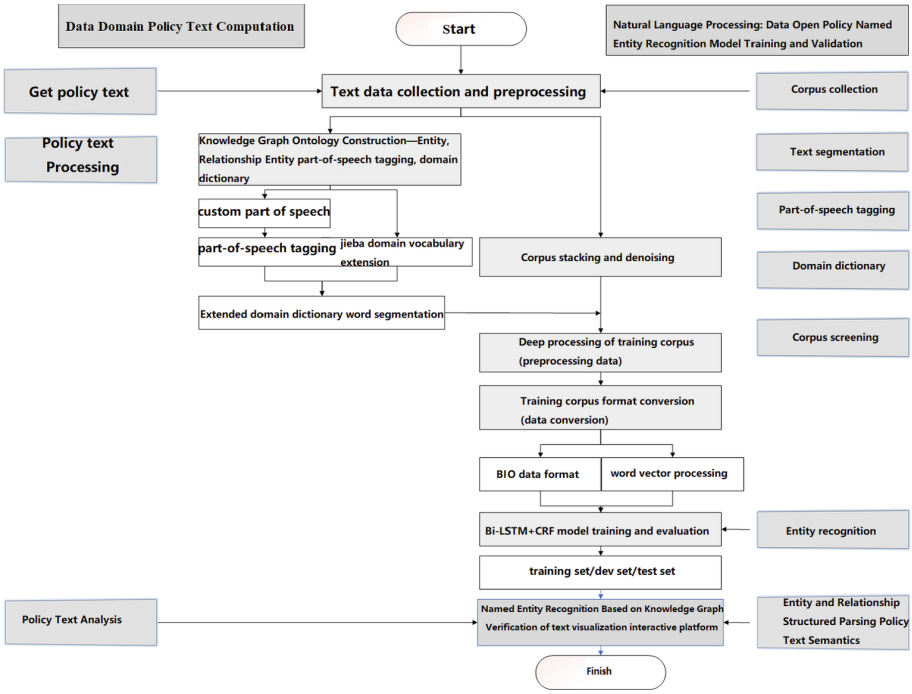

Fig. 1. Research framework of policy text analysis model

3.2 Ontology Construction of Policy Texts

According to the text analysis requirements of the data policy in this paper, a knowledge framework is constructed for the application scenarios: that is, according to the important principle of "opening while protecting" in the research hotspot of data development policy [16], the "open data" and "data open data" of the "data open" policy are customized. The two entity types of "safe" constitute a dictionary, and the entity types are set as: open and safe. The concepts of "open data" and "data security" based on the data open policy, and the knowledge graph ontology of the "collaboration" relationship are constructed, and the policy text is completed. Knowledge representation of corpus.

3.3 Data Collection and Preprocessing

This research selects text documents about data open policy in policy documents published by CNKI, government websites, and obtained on the Internet, and selects policy text corpus in the field of "data open" as a collection of policy text data, including policy text document corpus, government The website publishes the original text of the policy, the text of the policy interpretation in the network, the academic achievement literature on the policy interpretation of the research field in the CNKI literature knowledge base, and the self-built data open policy corpus as the data source [17], and completed 76 data open policy documents, There are 25 data security policy documents. After many

manual extractions, 591 pairs of "data open" keywords and corpus sentences, and 295 pairs of "data security" keywords and corpus sentences are extracted. Complete the part-of-speech tagging of the data open in the data-safe topic dictionary, and expand the Jieba word segmentation vocabulary Opendata. The BIO tag category is indexed for the filtered corpus sentence, which is used as the training input of the model algorithm after the policy text corpus is processed.

3.4 Build Key Technologies

This paper uses the Bi-LSTM + CRF [18] named entity algorithm to complete the entity extraction task of the knowledge graph of the data opening policy text. Complete each word category label of an entity in a given sentence. The general idea is as follows: First, use Word2vec for the low-dimensional and dense word vector matrix of the dictionary corpus and the domain dictionary of the part-of-speech to supply Bi-LSTM + CRF, and load the training set After and the test set, use Bi-LSTM to automatically extract features from the context information of the word (character) vector sequence in the input policy text, and then provide the CRF model as a feature to process the dependency information between tags, and select the most suitable one. Predict the tag sequence to complete named entity recognition. Second, predict the results, that is, after training the model, reload the model, input new prediction text, and identify the named entity in the policy text. First load the character dictionary, then load the model, then preprocess the input text into a character sequence, and then the model predicts the output entity category at each moment. Add and import data, according to the trained word vector, find the corresponding word vector through Word2vec, build the Bi-LSTM + CRF model, calculate the loss function, optimize the loss function, update the model parameters, test the model function, and adjust the data format to suit the Model input, evaluate the training effect of the model.

According to the characteristics of the domain policy, this research designs the training system architecture of the named entity recognition model for policy text, as shown in Fig. 2. The architecture consists of 3 layers: look-up layer, bidirectional LSTM layer and CRF layer. The model uses the BIO annotation set in the Bakeoff-3 evaluation. It is completed in three steps: feature representation, model training, and model classification. It inputs a set of corpus (character) vectors about open data and open policy, and outputs a set of predicted tag sequences. The following According to the three aspects of policy text feature representation, model training and model classification, the key issues of the training process are described.

Feature Representation

Before the neural network model is executed, the words of the input data open policy corpus need to be converted into data. First, take the sentence as a unit, use one-hot coding to form the word embedding vector layer in each data open and data security field in each sentence of the data open corpus. In the first look-up layer, use the pre-The training or randomly initialized embedding matrix maps each word in the sentence from a one-hot vector to a low-dimensional dense word vector (character embedding), that is, designing and building a neural network model and representing the symbolic

features of the text as distributed Feature information; Unlike traditional LSTM, Bi-LSTM considers both past features (extracted by forward process) and future features (extracted by backward process). The backward process is equivalent to inputting the original sequence into the LSTM in reverse. For example, the forward LSTM expresses the input sequence (x1, x2,…, xt,…, xn) as), and then uses the reverse LSTM to convert the input sequence (x1, x2,…, xt,…, xn) are expressed as (……,), and the concatenation of and is taken as xt as the final result.

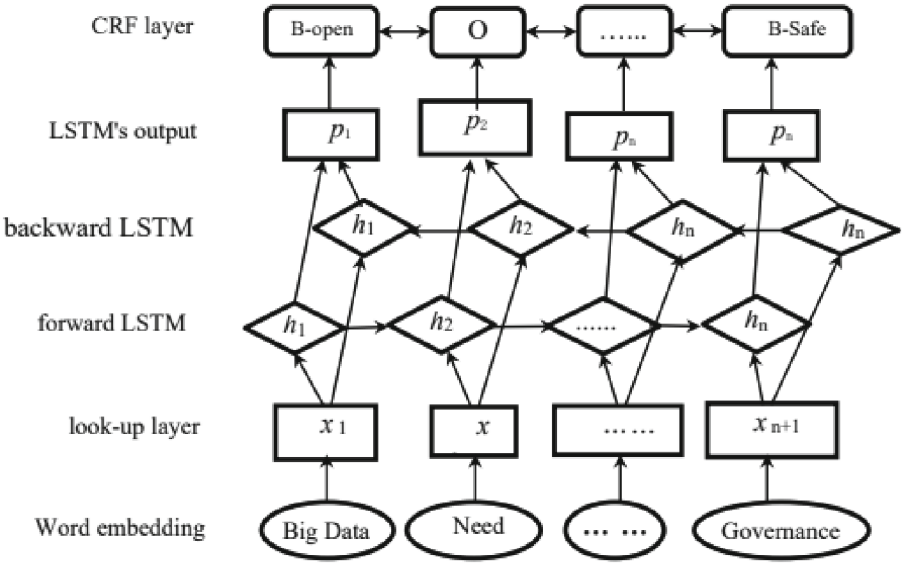

Fig. 2. Entity extraction model of data open policy text knowledge graph

Model Training

Use the splicing vector encoded by Bi-LSTM as the feature representation ht to perform softmax classification, obtain the label of each word, and splicing the K-dimensional vector (K is the number of labels) obtained from the representation of each word to obtain the input P, P is The n*k-dimensional matrix is uniformly input into the CRF model as a feature. The score of each sentence is shown in formula (1):

$$S(x, y) = \sum_{i=0}^{n} A_{y_i, y_j,} + \sum_{i=1}^{n} p_{i, y_j} \tag{1}$$

Among them, A is a transition matrix, which represents the probability of transition from one label to the next label in the label sequence. The parameters that need to be trained are: the parameters in Bi-LSTM and the transition probability matrix A in CRF, the supervised learning method is used in Bi-LSTM + CRF training, by maximizing the probability of predicting the real label sequence (take the logarithm of the probability and then take Negative, and then use gradient descent algorithm to optimize) to update

the parameters in Bi-LSTM and the transition probability matrix A in CRF. At the beginning of training, "the real label sequence will not correspond to the maximum probability value", but through continuous iterative optimization of the samples, "the real label sequence should correspond to the maximum probability value" will eventually be realized; when Bi-LSTM + CRF is tested, directly according to the training The good parameters are used to obtain the scores corresponding to all possible prediction sequences, and finally the prediction sequence corresponding to the maximum score is taken as the final prediction result.

Model Classification

Use the trained neural network model to classify policy texts. First, use Bi-LSTM to represent the input text, and then input it into the CRF to classify the sentences of data openness, fine-grained data openness and data security in the policy, and the classification labels are entity type and BIO three kinds of label combinations, and finally output the classification result, that is, determine the boundary between the part of speech and the word, so as to complete the named entity recognition of the policy analysis.

Model Evaluation

①Evaluation indicators

The three important indicators in the neural network evaluation system include the accuracy rate, the recall rate, and the harmonic average F1. The specific formulas (2), (3), and (4) are as follows:

$$\text{Precision (Precison, referred to as P value) } P = TP/(TP + FP) \tag{2}$$

$$\text{Recall rate (Recall, referred to as R value) } R = TP/(TP + FN) \tag{3}$$

$$\text{Harmonic mean } F1 = 2 * P * R/(P + R) \tag{4}$$

Among them, TP represents the number of correctly recognized entities by the policy text model, FP represents the number of incorrectly recognized entities, FN represents the number of unrecognized entities, and TP + FP represents the total number of recognized entities, that is, FP + FN is the name of the class Total number of entities. P represents the ratio of the number of correctly identified entities in the prediction results to all identified entities, and R represents the proportion of entities that are correctly identified. F1 represents the harmonic mean of the precision rate P and the recall rate R. When both P and R are high at the same time, a higher F1 value can be obtained.

②Evaluation effect

According to the training log information of train.log, the experimental results of policy text entity recognition are displayed, as shown in Table 1. Without adding artificial features, the deep learning model iterates 46 times, and the model in the dev set is saved. Its accuracy P value, recall rate R value and the corresponding evaluation index of the CRF + + machine learning algorithm in the F1 control experiment The comprehensive index and the average index are significant promote.

Table 1. Experimental results of entity recognition model in policy text

Algorithm	P value	R value	F1 value
Bi-LSTM + CRFmodel	82.76%	92.31%	87.27
CRF + + model	66.54%	75.00%	70.52

4 Application of Knowledge Graph in the Field of Policy Text

4.1 Visual Application Interaction Platform

In this study, the combination of Flask + Uwsgi is used to request the API data interface of the NRE (Bi-LSTM + CRF) model from the web application server to complete the policy text analysis application request. Vue2.0 + Echarts + elementUI realizes the interactive front-end display of the visual platform. This paper selects the data opening policy text of the "Outline of Action for Promoting the Development of Big Data" issued by the State Council in 2015 [19], as shown in Fig. 3.

Fig. 3. Data opening policy text input interface

Obviously, the visual interactive verification platform provides an intuitive and visible analysis basis for the application of policy analysis and exploration. As shown in Fig. 4, the direct result of entity identification of the policy text of the "Outline of Action for Promoting Big Data Development" identified 433 "data open" entity objects and 77 "data security" entity objects. It can be seen that after applying the knowledge graph knowledge table method and customizing the knowledge ontology modeling of the relationship between "data openness", "data security" and "synergy", the data opening policy text for the "Outline of Action for Promoting the Development of Big Data" can be completed. Automatic calculation.

Further, the distribution of entities and relationships in the graph is used to more clearly show the situational distribution of the collaborative relationship between "data openness" and "data security" in the data openness policy text of the "Outline of Action for Promoting the Development of Big Data", as shown in Fig. 5.

Fig. 4. Entity recognition results for data opening policy text

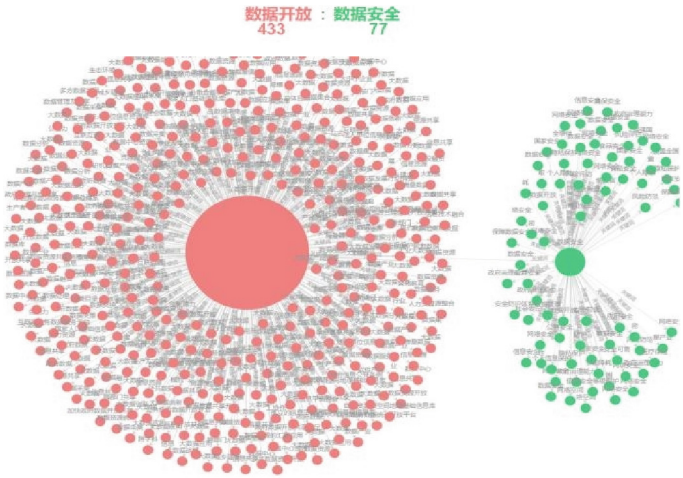

Fig. 5. Distribution of collaborative situation between data openness and data security

4.2 Collaborative Situation Analysis of Data Policy Texts

According to the analysis of the data opening policy in the "Outline of Action for Promoting the Development of Big Data", combined with the results of the visual map of the synergistic relationship between "data opening" and "data security", the analysis shows that: First, this document emphasizes the openness and sharing of big data while at the same time., to a certain extent, taking into account data security issues, and following the principle of "opening and sharing while paying attention to data protection". Second, from the analysis of the semantic ratio of "data" openness and "data security", as well as the degree of their synergistic relationship, it is clear that data openness and data security

show an uneven distribution. The guiding significance of open data sharing. The formulation and implementation process of various local data opening policies can combine different regional characteristics, application field characteristics, and implementation object characteristics to strengthen data security.

5 Conclusion

This paper applies the natural language processing technology of computer science and linguistics, takes the semantic intelligent analysis of data development policy texts as the research goal, and takes the "cooperative problem of opening while protecting" under the demand of data opening as the research application scenario. The graph represents the framework of policy collaboration knowledge in the field of data development. According to the ontology definition of the concepts of "data openness" and "data security" defined in the application scenario, the synergistic relationship between the two is used as the relationship definition, and the Bi-LSTM + CRF algorithm is selected for training data development. The policy text analysis model, after analyzing the data opening policy, obtains an accurate semantic layer analysis of a pair of synergistic relationships of "data openness" and "data security". This research also uses the data opening policy text of the "Outline of Action for Promoting Big Data Development" to analyze the collaborative situation of data development policy texts. From the perspective of knowledge extraction, this paper aims at the semantic analysis of data opening policy texts. There are still some limitations in the exploratory application of the method based on model training. Further exploration and research on knowledge relationship extraction concerned in policy texts is needed.

References

1. ACM DL.The federal big data research and development strategic plan[EB/OL]. (11 Aug 2021).https://dl.acm.org/citation.cfm?id=3027595
2. Industrial strategy:artificial intelligence sector deal[EB/OL]. (11 Aug 2021). https://www.gov.uk/government/publications/artificial-intelligence-sector-deal
3. Ma, H., Hong, W.: Pilot research on policy coordination of open government data in my country. Library Construction **4**, 61–68 (2018)
4. Li. J., Liu, Y., Huang, C., et al.: Reshaping policy text data analysis with bibliometric research: the origin, migration and method innovation of policy bibliometrics. J. Public Administration (2) (2015)
5. Pei, L., Sun, J., Zhou, Z.: Policy text calculation: a new way of interpreting policy text. Books Inf. **6**, 47–55 (2016)
6. Peng, Z., Gao, F.: Quantitative research on china's mineral resources security policy texts from the perspective of policy tools. J. Central South Univ. (Soc. Sci. Ed.) **27**(05), 11–24 (2021)
7. Huang, C., Ren, T., Li, J., et al.: Responsibilities and interests: a study on the evolution of intergovernmental cooperation in China's science and technology innovation policy based on quantitative analysis of policy literature. Manage. World **12**, 68–81 (2015)
8. Tian, J., Yang, Z.: Homogeneity and difference: a bibliometric analysis of the implementation policy of provincial government power list system. J. Intelligence **36**(5), 75–81 (2017)

9. Chu, D.: An analysis of rural green development policy texts: based on the dimensions of instrumentality and synergy. J. Zhengzhou Univ. (Philos, Soc. Sci. Edition), **54**(02), 14–21+126 (2021)
10. Yang, Z., Tian, J.: Policy bibliometric research on open utilization of government data: a three-dimensional analysis perspective. J. Intell. **37**(12), 175–181 (2018)
11. Jiang, X.: Research on open and sharing policy of scientific data of foreign funding institutions——analysis of policy text based on NVivo 12. Mod. Intell, **40**(08), 144–155 (2020)
12. Hong, W., Ma, H.: Research on the evolution and synergy of my country's open government data policy: Based on the analysis of policy texts from 2012 to 2020. J. Intell. **40**(10), 139 (2021)
13. Zhang, T., Ma, H.: Collaborative research on open data and data security policy based on policy text computing. Intell. Theory Practice, **43**(06), 149–155+141 (2020)
14. Mao. Z., Zheng, F., Huang, Y.: Research on government data opening from the perspective of policy coordination. Electronic Government Affairs, (9), 14–23 (2018)
15. Chen, X., Li, R.: Research on the hot Spots concerning my country's mass innovation and entrepreneurship policy based on Co-word analysis. Univ. Electron. Sci. Technol. China (Soc. Sci. Ed.) **21**(02), 9–17 (2019)
16. Catelli, R., Casola, V., Pietro, G.D., et al.: Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification. Knowl.-Based Syst. **213**(1), 106649 (2021)
17. Ma, H., Pu, P.: Analysis of the research status of open data policy at home and abroad and judgment of research trends in my country. J. Chin. Library **41**(5), 76–86 (2015)
18. Ma, H., Zhang, T.: Research on the construction of corpus for wisdom service from the perspective of literature information Inf. Theory Practice **6**, 124–130 (2019)
19. State Council of the People's Republic of China: Action Outline for Promoting the Development of Big Data. Group Technol. Moder. Production **32**(3), 8 (2015)