Research on entity extraction based on subway energy consumption patent

Jinbing Ha, Yongjie Shi *

School of Economics & Management, NJUST University, Nanjing, China

* Corresponding author: 928181574@qq.com

Abstract. In order to reduce operating costs and fulfill social responsibilities, enterprises and researchers in the field of rail transit need to study more efficient energy-saving technologies. Patent is one of the important ways to obtain information and knowledge related to technological innovation, but most of the existing patent analysis systems are statistical analysis of external patent information, and the results are not intuitive enough, which is not conducive to more in-depth mining of patent information. This paper introduces knowledge mapping related technologies to deconstruct and organize patent information to support scalable patent query and analysis requirements.

Keywords: Entity extraction, Knowledge map, Patent analysis.

1. Introduction

The subway is responsible for promoting the green transformation and upgrading of the city. Therefore, many cities continue to expand the operation scale of subway lines, resulting in an increase in total operating power consumption and operating costs, which is also a big burden on local finance [1]. Enterprises in the field of rail transit need to study better energy-saving technologies and tap the potential of energy conservation and carbon reduction in the rail transit industry. According to the statistics of the World Intellectual Property Organization, if patent data can be effectively used, enterprises can save 60% of the technology research and development time and about 40% of the research and development funds [2]. This research introduces the technical methods related to the knowledge map, studies the method of automatically extracting entities from a large number of patents, and builds a knowledge map of subway energy consumption based on patent texts in the field, so as to effectively organize and manage patent knowledge in the field of subway energy consumption, tap the internal relevance of patent knowledge, and further promote the efficiency of energy conservation technology research and development in the field of rail transit.

2. Research status of entity extraction

Named Entity Recognition, refers to identifying the location of the representative entity from the text, extracting entity information, and classifying it into a specific type [3]. At present, the method of entity extraction is the method of deep learning. Collabert et al. first proposed the NER method based on neural network [4]. Then, deep learning entity extraction methods based on recurrent neural networks and Long Short-Term Memory (LSTM) emerged [5]. Huang et al. built CRF layer on LSTM model, obtained text context information through LSTM, and used CRF to obtain tag sequence information [6]. Ji et al. proposed using similar information retrieval related measures to select documents, and improved the existing named entity classifier to solve the problem of unlabeled data selection [7].

Most of the patent related knowledge maps are based on the external characteristics of patents, supplemented by subject words, etc. The knowledge maps show the relationship between patents and patents, and do not dig into the content of patent literature in a fine-grained way [8]. In recent years, some scholars believe that the technical atlas can be drawn based on the text representation model of patent documents [9], but it is also relatively simple to build the atlas through code frequency, which is still mainly based on the external characteristics of patent documents. Ma Guobin went deep into the text content of patent documents to build a knowledge map, and refined patents from documents

to smaller structured knowledge granularity to facilitate refined knowledge management and application [10].

In general, although the construction of knowledge maps using patent classification numbers and subject words is relatively related to the content of patent literature, there is no fine-grained analysis of the content of patents. Only from the field scope, patents are considered innovative. From this perspective, there is still much research space for the application of patent knowledge maps.

3. Entity extraction method of subway energy consumption patent abstract

3.1. Semantic model construction of patent abstract

At present, the semantic classification of patent texts is diversified, and there is no unified standard. Based on the existing research status of abstract entity classification and the text characteristics of subway energy consumption patent abstract, this paper defines eleven entity types in the abstract, which conforms to the semantic expression in most relevant patent abstracts. Entity type definition is shown in Table 1.

Table 1. Entity Type Definition

Numble	Entity Type	Explain	Numble	Entity Type	Explain
1	Domain	Application object or technical field	6	Algorithm	Use computer description to solve problems
2	Equipment	Articles without original functions after splitting	7	Part	Components of equipment
3	Method	Means taken to achieve an end	8	Data	Various data and parameters
4	System	A whole with some functions of	9	Physical	Various physical terms
5	Materials	Raw materials used to make articles	10	Shape	The manifestation of things or substances

3.2. Entity Dataset Construction

The patent data used in this paper is from the patent database of the State Intellectual Property Office. According to the defined semantic model, 500 pieces of the collected patent data are randomly selected according to the proportion of IPC classification numbers for physical manual annotation. After determining the dataset to be labeled, it is necessary to manually label the dataset. The annotation strategy is that two annotation personnel annotate the same data first, and the same data forms two annotation files, which are then submitted to the administrator for verification and correction, so as to minimize manual annotation errors, improve the quality of annotated entity data sets, and lay a good foundation for the construction of relational data sets. Details of data scale are shown in Table 2

Table 2. Entity dimension data size statistics

Patent Part Number	Patent Number	Sentence Number
B	145	349
F	95	253
H	85	204
E	70	183
G	75	158
C	30	81
Total	500	1228

3.3. Entity extraction model

The task of entity recognition is to automatically recognize the words with specific meaning in the patent abstract text by using the algorithm model. The Bert-BiLSTM-CRF model is adopted in this paper, and the model structure is shown in Figure 1.

Get word vector sequence $X = (\overrightarrow{x_1}, \overrightarrow{x_2} \dots, \overrightarrow{x_n})$ through Bert model after inputting text. BiLSTM model conducts bidirectional training on it and outputs feature vector sequence $Y = (\overrightarrow{y_1}, \overrightarrow{y_2} \dots, \overrightarrow{y_n})$, Where eigenvector $\overrightarrow{y_i} = (\overrightarrow{y_{i1}}, \overrightarrow{y_{i2}} \dots, \overrightarrow{y_{in}})$ Indicates the probability distribution of the x_i tag of the i character in the input sentence. Finally, the CRF model is used to obtain the optimal tag annotation sequence.

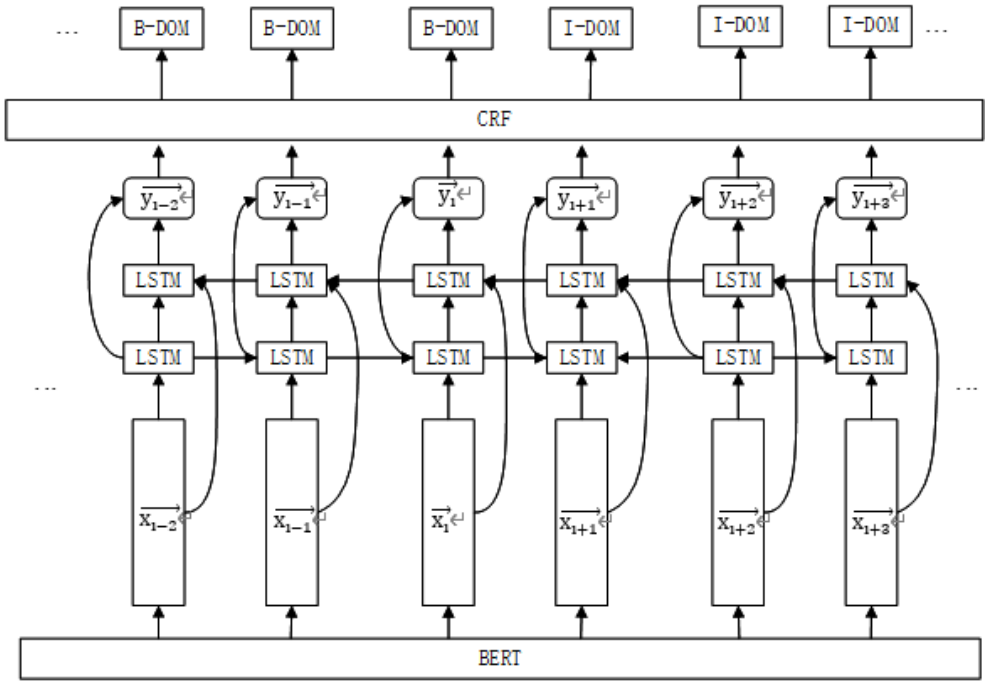

Figure 1. Bert-BiLSTM-CRF model structure

4. Experimental results and results analysis

4.1. Experimental parameter setting

The entity extraction experimental data is the subway energy consumption patent entity data set constructed by manual annotation. The data set was randomly divided into training set and test set at a ratio of 7:3. The settings adopted in this experiment: the maximum sentence length "Max_seq_length" during training is 300, the training "epoch" is set to 30, the learning rate "Learning_rate" is set to 0.0001, and the size of each batch of samples "Batch_size" is set to 32.

4.2. Experimental results and analysis

In the entity extraction task, in order to verify that each module of Bert-BiLSTM-CRF model used in this paper can play a positive role in entity recognition, this paper selects three groups of models, namely, BiLSTM, BiLSTM-CRF, Bert-CRF and Bert-BiLSTM, for comparative experiments. BiLSTM and BiLSTM-CRF all use word2vec static word vector to replace BERT to complete vector training. The experimental results are shown in Table 3.

Table 3. Comparison of experimental results

Model	Accuracy P	Recall R	F1 value
BiLSTM	76%	81%	78%
BiLSTM-CRF	78%	83%	81%
Bert-CRF	83%	86%	85%
Bert-BiLSTM-CRF	84%	90%	87%

It can be seen from the experimental results that CRF can learn the constraint relationship between tags, so it can effectively improve the model recognition effect. By comparing the BiLSTM-CRF model with the Bert-BiLSTM-CRF model, the accuracy, recall and F1 values of the word vector obtained by replacing word2vec with the Bert pre training model have increased significantly, indicating that the Bert pre training model has learned rich semantic expressions on large-scale corpus data sets, and can well understand the semantic information of texts in specific fields through fine-tuning. Bert-BiLSTM-CRF model and Bert-CRF model have similar results. After the introduction of BiLSTM model, the extraction effect has been improved to a small extent. Therefore, Bert-CRF model can be considered when computing resources are limited. In general, when the computing resources are sufficient, the BERT-BiLSTM-CRF model has a better recognition effect on the entity recognition task of subway energy consumption patent text.

4.3. Construction of patent knowledge map of subway energy consumption

The knowledge map is a structured semantic knowledge base, which describes the association between entities in the form of a graph, in the form of (entity-relationship-entity). In this paper, the abstracted entities are linked to the patent information itself, and finally the extracted data is stored in the Neo4j map database, as shown in Figure 2. Through the knowledge map constructed, the specific information of subway energy consumption patents is aggregated and organized, and the knowledge map based on subway energy consumption patent data is constructed, providing users with fine-grained and structured knowledge representation and services.

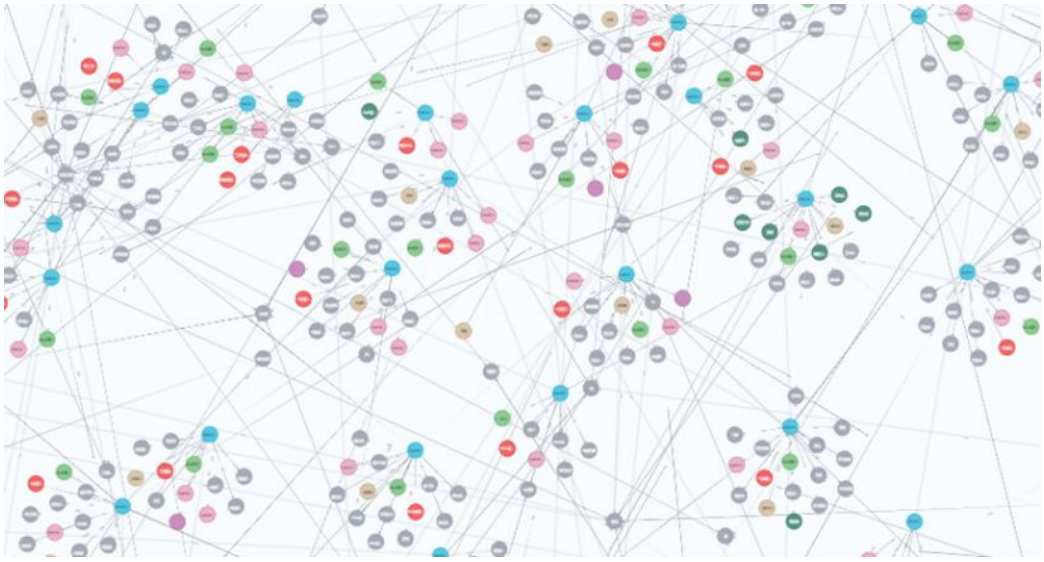

Figure 2. Patent Knowledge Map of Metro Energy Consumption (Part)

5. Summary

In this paper, the semantic model of metro energy consumption patent abstract entity is established based on the external characteristics of patent and the specific knowledge in the text, aiming at the knowledge requirements in the field, and the structural expression of important technical information of patent is solved. In the process of pattern layer construction, all entity categories required by entity knowledge extraction task are defined in detail, so that they can adapt to the knowledge expression

patterns in most subway energy consumption related patents. Build a patent knowledge map of subway energy consumption, and display the query results in a visual form, thus helping to improve the patent utilization efficiency of practitioners.

Acknowledgements

The authors gratefully acknowledge the financial support from Postgraduate Research & Practice Innovation Program of Jiangsu Province funds.

References

[1] 2021 Annual Statistics and Analysis Report of Urban Rail Transit [J]. Urban Rail Transit, 2022 (07): 10 - 15. DOI: 10.14052/j.cnki.china.metros.2022.07.012.

[2] LI Miao, YU Jiaxin, YIN Zhijuan, et al. Research on technological innovation opportunities based on patent analysis and technological evolution[C]//2021 2nd International Conference on Intelligent Design (ICID). IEEE, 2021: 466 - 471.

[3] Mohit B. Named entity recognition[M]. Natural language processing of semitic languages. Springer, Berlin, Heidelberg, 2014: 221 - 245.

[4] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (almost) from Scratch [J]. Journal of Machine Learning Research, 2011,12 (1): 2493 – 2537.

[5] Hochreiter S J, Schmidhuber R A. Long short–term memory[J]. Neural Computation, 1997, 9 (8): 17305 – 1780.

[6] Huang Z, Xu W, Yu K. Bidirectional LSTM–CRF models for sequence tagging [J]. arXiv preprint arXiv:1508.01991, 2015.

[7] Ji H, Grishman R. Data selection in semi–supervised learning for name tagging [C]. Proceedings of the Workshop on Information Extraction beyond the Document, Sydney, Australia, July 2006: 48 – 55.

[8] Cao Shujin, Li Ruijing. Construction and application of innovation knowledge map based on patent literature abstract [J/OL]. Information theory and practice https://kns.cnki.net/kcms/detail/11.1762.G3.20220929.1204.004.html

[9] Pan Donghua, Xu Keke. Research on the drawing method of technical knowledge map based on patent document classification code [J] Journal of Information Science, 2015, 34 (8): 866 - 874.

[10] Ma Guobin. Research on Patent Knowledge Retrieval Based on Knowledge Graph [D] Harbin: Harbin University of Technology, 2021.