Intelligent Technology Assessment of High-Speed Railway Based on Knowledge Graphs

Chenchen Liu [1, a, *], Hongwei Wang [2, b], and Lin Wang [1, c]

[1] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China;

[2] National Research Center of Railway Safety Assessment, Beijing Jiaotong University, Beijing, China.

[a, *] 22120245@bjtu.edu.cn, [b] hwwang@bjtu.edu.cn, [c] 21120249@bjtu.edu.cn

Abstract. This paper introduces an intelligent technology assessment framework for high-speed railways based on a knowledge graph approach. We employ rule-based knowledge extraction algorithms and a Bert-BiLSTM-CRF model for entity extraction from technical texts. Subsequently, we establish relationships among various entities, constructing a knowledge graph specific to high-speed railways. The knowledge graph is stored in a Neo4j graph database in triple format. Furthermore, we establish a comprehensive evaluation metric system, integrating knowledge graph insights to assess the utility of enabling technologies. We employ the CRITIC weighting method to calculate the value assessment results for target technologies. Simulation results indicate that the training results for word segmentation and sentence splitting are favorable, and the Bert-BiLSTM-CRF model achieves convergence in accuracy, recall, and F1 score after 30 iterations. The technical assessment results in this paper align closely with the actual technological value assessment.

Keywords: knowledge graph; high-speed railroad; technology assessment; Bert; BiLSTM-CRF.

1. Introduction

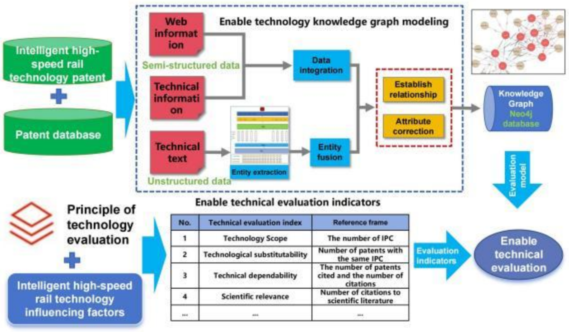
There remains a lack of systematic organization and summarization in several aspects of intelligent high-speed railways, including the construction of an enabling technology framework, the interplay between theory and technology, and the absence of practical modeling and assessment methodologies. With the impetus provided by intelligent technology, knowledge graph application techniques have come to the forefront.

Some scholars have undertaken preliminary explorations and attempts in the domain of railway safety knowledge graphs. Liu [1] applied knowledge graphs to the analysis of railway operational accidents, revealing potential patterns in accidents by describing incidents and hazards within a heterogeneous network. Wu [2] introduced a method using BiLSTM and CRF [3] to extract CTCS-3 knowledge from unstructured data, generating entity relationship triplets to aid machines in comprehending CTCS-3 knowledge efficiently and specifically.In the realm of technical assessment research, Zhu [4] introduced patent citation quality as a measure to assess the quality of invention patents and validated its effectiveness and necessity within the context of the new energy industry. Gu [5] established a comprehensive indicator system based on three aspects: technical benefits, scope of protection, and technical content. This framework provides a reference for accurately evaluating technology quality. Lee [6] proposed a soft computing-based approach to technical assessment, employing fuzzy set theory and genetic algorithms to implement a patent technology evaluation mechanism.

2. System Model

A knowledge graph is a knowledge representation method based on graph theory and semantic networks, which enables the unified representation and management of knowledge and data from various subsystems, facilitating knowledge sharing in intelligent high-speed railways. The technology assessment framework for intelligent high-speed railway enabling technologies based on knowledge graphs is illustrated in Figure 1.

Simultaneously, principles for constructing technical evaluation indicators are established, and factors influencing technical value are analyzed. Based on these principles, an evaluation indicator system for intelligent high-speed railway enabling technologies is developed, considering the completeness, accessibility, and objectivity of technical indicator data. Finally, leveraging the enabling technology knowledge graph, an evaluation of the intelligent high-speed railway enabling technology system is performed through the analysis of various enabling technology assessment indicators.


Intelligent high-speed rail technology patent
Patent database
Enable technology knowledge graph modeling
Web informat ion
Semi-structured data
Technical informati on
Technical text
Unstructured data
Entity extraction
Data integration
Entity fusion
Establish relationship
Attribute correction
Knowledge Graph Neo4j database
Evaluation model
Principle of technology evaluation
Intelligent high-speed rail technology influencing factors
Enable technical evaluation indicators
Evaluation indicators
Enable technical evaluation

No.	Technical evaluation index	Reference frame
1	Technology Scope	The number of IPC
2	Technological substitutability	Number of patents with the same IPC
3	Technical dependability	The number of patents cited and the number of citations
4	Scientific relevance	Number of citations to scientific literature
...	...	...

Figure 1. Knowledge graph-based evaluation of intelligent high-speed railway enabling technology system

2.1 Enabling Technology Knowledge Graph Construction

2.1.1 *Knowledge Extraction:* Knowledge extraction mainly includes two tasks of named entity recognition and relationship extraction. In view of the multi-source and heterogeneous characteristics of the intelligent high-speed rail enabling technology and its patent information data, in order to improve the data accuracy and the completeness of the knowledge graph, the rule-based and deep learningbased methods are comprehensively applied to complete the knowledge extraction work. For semi-structured web data, custom rules are used for entity recognition. For unstructured text data, such as patent text, the Bert-BiLSTM-CRF model is constructed to realize automatic knowledge extraction. Finally, the extraction results of the two parts of the collection are stored into the Neo4j graph database.

2.1.2 *Knowledge Storage:* In this paper, structured data is stored using Neo4j. The essence of knowledge graph storage lies in the storage of triples, where each triple consists of a subject, predicate, and object.

2.2 Enabling Technology Assessment Indicator System Construction

The technical value is crucial in establishing an enabling framework for the assessment of technology. Upon the completion of the knowledge graph for intelligent high-speed railways technology, data values for various indicators can be computed. Following the methodology outlined in reference [7], twelve specific indicators have been selected. Their respective names and reference data are presented in Table 1. We employ the CRITIC weight method to calculate the weights of these indicators.

Table 1. Patent Technology Value Assessment Index System

Evaluation indicators	**References**	**CRITIC weight[7]**
Technical coverage	Number of IPC international classification numbers	0.0641
Number of claims	Number of claims in the claim	0.649
Technical	Number of patents with the same IPC number	0.0744

Evaluation indicators	**References**	**CRITIC weight[7]**
substitutability		
Number of homologous patents	Number of patents in the sametechnical field or with similar subject matter keywords	0.0667
Technology dependency	Number of patent citations, Number of references cited in the patent	0.0876
Scientific relevance	Number of references cited in the patent	0.0845
Impact index	Number of patent citations	0.1049
Technology life cycle	Growth rate of the number of homologous patents	0.0782
Remaining validity period	Remaining validity of the patent	0.0776
Patent efficiency	Number of active patents, number of granted patents	0.1013
Patent grant rate	Number of granted patents, number of homologous patents	0.1056
Technical Stability	Legal status of patents	0.0902

On this basis, this paper defines the technical value scoring formula as follows:

$$Y_j^{'} = \frac{max(Y_j)-Y_j}{max(Y_j)-min(Y_j)}(1)$$

$$Score = \sum_{j=1}^{12} W_j \cdot Y_j^{'}(2)$$

Where Yi denotes the data value of the ith indicator, and Score denotes the scoring result of the target technology value, which means that the value of each indicator is multiplied by its corresponding weight value and then summed up.

3. Knowledge Extraction Algorithm

Semi-structured data is processed using rule-based knowledge extraction methods. Unstructured patent text data requires the use of knowledge extraction algorithms based on the Bert-BiLSTM-CRF model.

3.1 Rule-based knowledge extraction

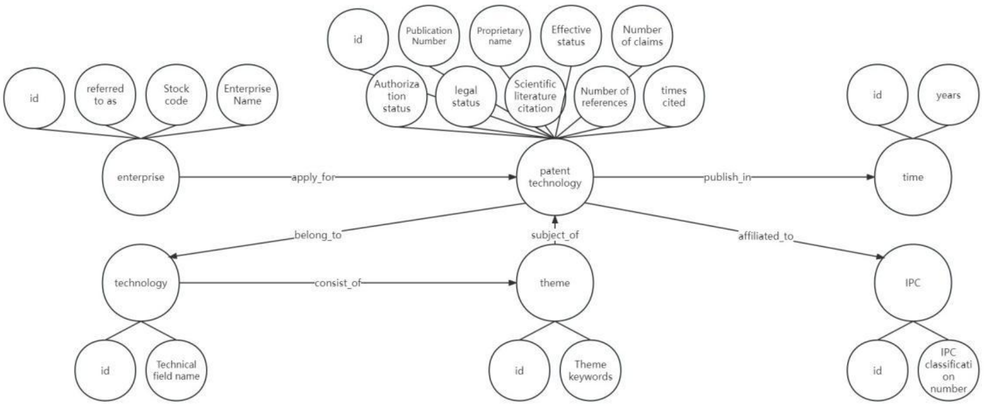


Figure 2. Knowledge graph framework

The fundamental concept behind rule-based knowledge extraction is to manually identify potential keywords or phrases that may indicate relationships. Subsequently, sentences containing relationship information are extracted based on these keywords or phrases within the text. Finally, specific patterns are employed through template matching to extract triples that conform to predefined patterns. In the field of intelligent high-speed railway enabling technology patents, entity types often include patent publication numbers, years, months, technology domain names, thematic keywords, IPC classification codes, and more. After entity extraction is completed, the data is transferred to the Neo4j graph database using py2neo, and an intelligent high-speed railway enabling technology knowledge graph is constructed, as illustrated in Figure 2.

3.2 Maintaining the Integrity of the Specifications

BERT(Bidirectional Encoder Representations from Transformers)is a pre-trained language model, as illustrated in Figure 3. Its model architecture consists of multiple layers of Transformer Encoders, with each layer comprising two sub-layers: Self-Attention and Feed-Forward. To be more specific, BERT incorporates the Multi-Head Attention mechanism from the Transformer Encoder. This mechanism calculates self-attention across multiple heads for the input sequence, thereby obtaining contextually relevant representations.

The core idea behind BILSTM-CRF is to perform sequence labeling by combining BiLSTM (Bidirectional Long ShortTerm Memory networks) and CRF (Conditional Random Fields). On the other hand, is used to model dependencies between labels, thereby enhancing the accuracy of label predictions.

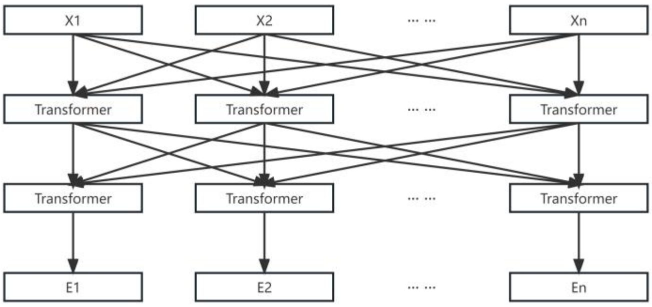


Figure 3. BERT model structure

The BiLSTM-CRF model is depicted in Figure 4.

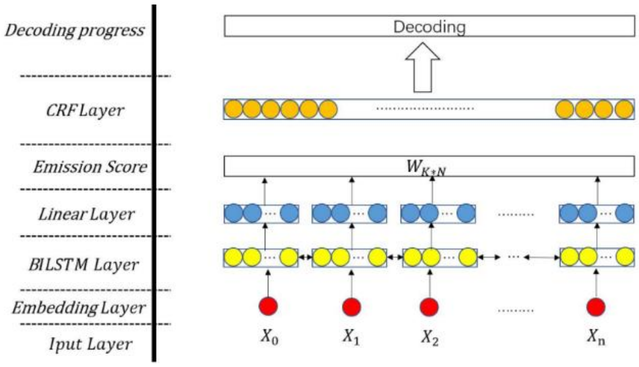


Figure 4. BiLSTM-CRF model structure

1) *Input Layer*: This layer corresponds to the high-speed railway intelligent technology text information that has undergone preprocessing with BERT.

2) *Embedding Layer*: This layer encompasses the process of converting sentences into word vectors.

3) *BiLSTM Layer*: The Bidirectional Long Short-Term Memory network is a widely used RNN (Recurrent Neural Network) model. The calculation for the hidden state h_t^f of the forward LSTM unit is as follows:

$$i_t^f = \sigma\left(W_{ix}x_t + W_{ih}h_{t-1}^f + b_i\right)(3)$$

$$f_t^f = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1}^f + b_f\right)(4)$$

$$o_t^f = \sigma\left(W_{ox}x_t + W_{oh}h_{t-1}^f + b_o\right)(5)$$

$$\tilde{c}_t^f = tanh\left(W_{cx}x_t + W_{ch}h_{t-1}^f + b_c\right)(6)$$

$$c_t^f = f_t^f c_{t-1}^f + i_t^f \tilde{c}_t^f(7)$$

$$h_t^f = o_t^f tanh\left(c_t^f\right)(8)$$

In this context, $\mathrm{i}_\mathrm{t}^\mathrm{f}$, $\mathrm{f}_\mathrm{t}^\mathrm{f}$, $\mathrm{o}_\mathrm{t}^\mathrm{f}$, and $\tilde{\mathrm{c}}_\mathrm{t}^\mathrm{f}$ respectively represent the input gate, forget gate, output gate, and new memory cell. σ denotes the sigmoid function, while W and b are the model's weight parameters and bias parameters. The calculation of the hidden state h_t^b of the backward LSTM unit is similar to that of h_t^f . The final output h_t is formed by concatenating the hidden states from both the forward and backward directions:

$$h_t = \left[h_t^f \cdot h_t^b\right](9)$$

4) The output of the Linear Layer is used for the probability distribution of labeling sequences, representing scores for each label sequence. Its shape is *[n, k]*, where *n* is the length of the input sequence, and *k* is the number of labels.

5) The CRF Layer's purpose is to find the most probable and optimal path among all possible paths. Assuming an input $x = [x_o, x_1, x_2, ..., x_n]$, where each x_i corresponds to a label vector with a dimension equal to the number of labels, the goal is to decode the corresponding label sequence $y = [y_o, y_1, y_2, ..., y_n]$ with a conditional probability formula as follows:

$$(y|x) = P\left(y_o, y_1, y_2, ..., \quad , \middle| x_o, x_1, x_2, ..., x_n\right)(10)$$

Assuming there are k labels and the text length is n, there will be $\mathrm{N} = \mathrm{k}^\mathrm{n}$ paths. If we represent the score of the ith path as S_i, the probability of a label sequence appearing is calculated as follows:

$$P(S_i) = \frac{e^{S_i}}{\sum_J^N e^{S_j}}(11)$$

The probability of the real path, represented as S_real, occurring is given by:

$$P(S_{real}) = \frac{e^{S_{real}}}{\sum_J^N e^{S_j}}(12)$$

The goal of the CRF layer is to continually increase the probability $\mathrm{P}(\mathrm{S}_\mathrm{i})$ of the real path. By maximizing loss, we aim to minimize the loss function. For ease of computation, the loss function is solved in log space, as follows:

$$\begin{aligned} loss &=- log\frac{e^{S_{real}}}{\sum_J^N e^{S_j}} \\ &=- \left(log\left(e^{S_{real}}\right) - log\left(\sum_J^N e^{S_j}\right)\right) \qquad (13) \\ &= log\left(e^{S_1} + e^{S_2} + e^{S_3} + ... + e^{S_n}\right) - S_{real} \end{aligned}$$

The above describes the iterative training process of the Bert-BiLSTM-CRF neural network, outlining the principles and workflow used for named entity recognition.

4. Numerical Results and Discussion

4.1 Sentence Breaking Training Results

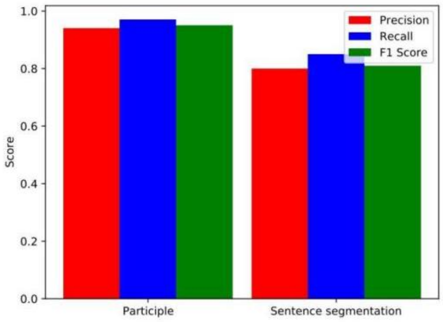


Figure 5. Sentence breaking training results

For the evaluation of segmentation and sentence boundary detection, the selected metrics include Precision, Recall, and F1 Score. Precision is calculated as Precision=TP/(TP +FP), where TP represents the number of correctly annotated positive instances, and FP represents the number of instances incorrectly labeled as positive by the model. Recall is calculated as Recall = TP/(TP + FN), with TP denoting the number of correctly annotated positive instances and FN representing the number of instances that were not correctly labeled as positive by the model. The F1 Score is the harmonic mean of precision and recall, calculated as F1=2 * (Precision * Recall) / (Precision+Recall).The experimental results, as depicted in Figure 5, indicate that the scores for both segmentation and sentence boundary detection are consistently above 0.8, demonstrating satisfactory training outcomes.

4.2 Bert-BiLSTM-CRF Model Training Results

As depicted in Figure 6, the Bert-BiLSTM-CRF model incorporated neural network training during its training process. The model encountered local optima, but through continuous iterations during the training process, it eventually converged, and the loss function reached stability. Subsequently, various performance metrics also gradually stabilized and reached optimal values. Precision, recall, and F1 score converged to around 85 after the 30th iteration.

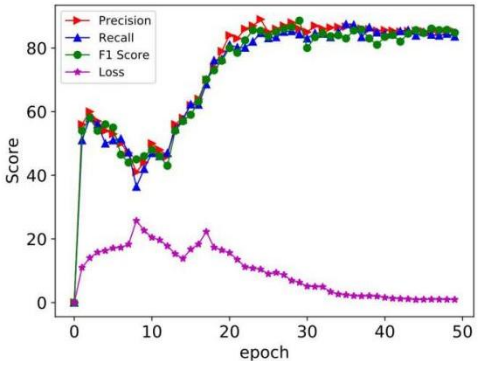


Figure 6. Bert-BiLSTM-CRF model Precision, Recall, F1-scoreLoss changes

4.3 Results of Technical Assessment

To assess enabling technologies, this paper parses the target technical texts and calculates the similarity between the entities extracted from the texts and the nodes in the knowledge graph. Subsequently, the knowledge graph is queried, and based on the retrieved information, values are assigned to various evaluation metrics. Finally, the value assessment of the target technology is computed using the CRITIC weighting method. In this section, ten patent technologies from the sample are chosen as validation subjects. Their corresponding scores are determined based on the assessment criteria for technical value, and the comparative results are presented as follows.

The comparative results between the technical value scores and the technical value degrees are presented in Figure 7. The red line represents the technical value scores obtained in this paper, while the blue line represents the technical value degrees used for comparison. From the figure, the score trends for all patent technologies are consistent.

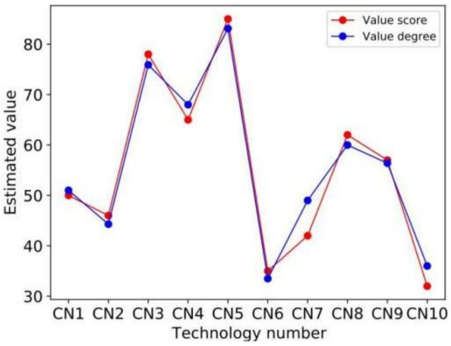


Figure 7. Technology value score comparison

5. Conclusions

This paper provides an overview of the current research landscape in knowledge graphs and technology assessment. It constructs a knowledge graph for high-speed railway technology using rule-based knowledge extraction algorithms and BiLSTM-CRF model-based knowledge extraction algorithms. Building upon this foundation, a knowledge graph-based approach for assessing intelligent high-speed railway technology is proposed. In future research, further exploration of its applications and refinement of algorithm models can be pursued, along with efforts to strengthen the integration and sharing of knowledge graphs with other application systems.

References

[1] J. Liu, F. Schmid, K. Li, and W. Zheng, “A knowledge graph-based approach for exploring railway operational accidents,” Reliability Engineering & System Safety, vol. 207, p. 107352, 2021.

[2] H. Wu, S. Li, H. Li, and Y. Wang, “A method for ctcs-3 knowledge extraction of unstructured data,” pp. 67 – 74, 2021.

[3] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” 2001.

[4] Z. Erdogan, S. Altuntas, and T. Dereli, “Predicting patent quality based on machine learning approach,” IEEE Transactions on Engineering Management, 2022.

[5] G. Li, H. Xue, Y. Wei-chun, and H. Chen, “A study on establishment of the patent application quality evaluation index system,” in 2018 Portland International Conference on Management of Engineering and Technology (PICMET). IEEE, 2018, pp. 1 – 6.

[6] C.-S. Lee, M.-H. Wang, Y.-C. Hsiao, and B.-H. Tsai, “Ontologybased gfml agent for patent technology requirement evaluation and recommendation,” Ph.D. dissertation, 2019.

[7] Y. Y. Meng, “Research on patent technology value evaluation method based on knowledge,” Ph.D. dissertation, 2017.