Economics and Public Policy
ISSN Print: 3104-4387
ISSN Online: 3104-4395
CD ISSN: 3104-4409
https://doi.org/10.63313/EPP.9009

A Study on Large-Scale Model Knowledge Extraction Methods for Enterprise-Level Data Governance

Yankun Li*. Meng Wu[a]. Longfei Ren[b]. Zhiheng Guo[c] and Wei Zhang[d]
China United Network Communications Corporation Limited Software Research Institute, Beijing, 100176, China
Email: *liyk71@chinaunicom.cn. [a]wum118@chinaunicom.cn. [b]renlf10@chinaunicom.cn.
[c]guozh110@chinaunicom.cn. [d]zhangw135@chinaunicom.cn

How to cite this paper: Li, Y. K., Wu, M., Ren, L. F., Guo, Z. H., & Zhang, W. (2025). A Study on Large-Scale Model Knowledge Extraction Methods for Enterprise-Level Data Governance. *Economics and Public Policy*, 1(2), 8-15. ISSN Print: 3104-4387; ISSN Online: 3104-4395.
https://doi.org/10.63313/EPP.9009
Published: 2025-12-22

Abstract

Enterprise data governance demands precise and standardized approaches to knowledge management. Addressing the limitations of traditional knowledge extraction methods when processing enterprise data governance documents, this study proposes an improved BERT-BiLSTM-CRF-based knowledge extraction method. It designs and implements a multi-level knowledge base architecture with an incremental update mechanism. A knowledge application service framework tailored for data governance is constructed, enabling intelligent decision support in core scenarios such as data standard formulation and quality diagnosis. This provides an effective technical solution for enterprise data governance.

Keywords

Data governance; Knowledge extraction; Knowledge graph; Deep learning; BERT-BiLSTM-CRF

1. Introduction

In the era of big data, the complexity and specialization of enterprise data governance continue to rise, placing higher demands on knowledge management capabilities. Existing knowledge extraction methods exhibit limitations such as insufficient semantic understanding and shallow knowledge association mining when processing specialized documents in the data governance domain. This paper proposes a knowledge extraction method based on an improved BERT-BiLSTM-CRF framework. It designs a multi-level knowledge base architecture and dynamic update mechanism tailored for data governance, enabling intelligent knowledge management and application. The research encompasses data governance knowledge feature analysis, knowledge extraction method design, knowledge base construction,

and application studies, providing technical support for enhancing enterprise data governance capabilities.

2. Design of Large-Scale Model Knowledge Extraction Methods for Data Governance

2.1. Analysis of Data Governance Knowledge Characteristics

Knowledge in the data governance domain exhibits characteristics of multi-source heterogeneity, complex hierarchical structures, and dynamic evolution. Analysis of 150,000 data governance documents accumulated over three years by a large enterprise reveals that data governance knowledge primarily spans four dimensions: data standards (32.5%), data quality (28.7%), metadata (21.3%), and data security (17.5%). Knowledge forms encompass structured rules (45.6%), semi-structured documents (35.2%), and unstructured text (19.2%). LDA topic modeling for document clustering revealed strong inter-knowledge correlations, with an average topic similarity of 0.72. By constructing a knowledge feature vector matrix, knowledge attributes were quantified through metrics including timeliness (timestamp), reliability (source rating), and relevance (co-occurrence frequency), laying the foundation for subsequent knowledge extraction[1]. As shown in Table 1, distinct feature differences exist across different knowledge types.

Table 1. Statistical Features of Data Governance Knowledge

Knowledge Type	Timeliness (Months)	Reliability (1-5)	Relevance (%)
Data Standards	12	4.8	82
Data Quality	6	4.2	75
Metadata	18	4.5	68
Data Security	3	4.9	71

2.2. Large-Model-Based Knowledge Extraction Framework

This study designed a hierarchical knowledge extraction framework based on pre-trained language models. The bottom layer employs the BERT model for text representation learning, the middle layer uses a BiLSTM-CRF model to identify entities and relationships, and the top layer applies a knowledge reasoning module to construct a knowledge graph. During the BERT pre-training phase, domain adaptation training was conducted using 2 million internal enterprise data governance documents, achieving a vocabulary expansion rate of 15%. For entity recognition, 35 fine-grained entity labels—including data objects, rule definitions, and quality metrics—were designed to address domain-specific characteristics[2]. Relationship extraction employed distant supervision, automatically generating training data from existing knowledge bases supplemented by 5,000 manually annotated high-quality samples as a validation set. The knowledge reasoning module, based on the TransE model, discovers latent knowledge through low-dimensional vector representations

of entities and relationships. Experimental results demonstrate that this framework effectively addresses complex knowledge extraction demands in enterprise data governance scenarios.

2.3. Core Algorithm Design

An improved deep learning algorithm was designed to address the characteristics of knowledge extraction in data governance. For entity recognition, an attention-based BiLSTM-CRF model was proposed, incorporating domain dictionary features to enhance recognition accuracy. The model structure comprises a word embedding layer, BiLSTM encoding layer, attention layer, and CRF output layer. Attention weights are calculated using Equation (1):

$$\alpha(t) = \text{softmax}\left(v^T \tanh(W_1 h_t + W_2 c)\right) \quad (1)$$

where h_t denotes the BiLSTM hidden state, c represents the context vector, and W_1, W_2, and v are trainable parameters. For relation extraction, a graph neural network-based approach constructs dependency syntactic trees between entities, utilizing R-GCN for information propagation. Each R-GCN layer update follows Equation (2):

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)}\right) \quad (2)$$

where N_i^r denotes the set of neighboring nodes connected to node i via relation r, and $c_{i,r}$ is the normalization constant[3]. This algorithm achieves superior performance compared to traditional methods on standard test datasets.

2.4. Knowledge Validation and Optimization Methods

To ensure the accuracy and reliability of extracted knowledge, a multi-level validation and optimization mechanism was established. First, based on the structural characteristics of knowledge graphs, a rule-based validation system was designed, encompassing checks for entity completeness, relationship validity, and attribute consistency[4]. Second, leveraging knowledge graph embedding techniques, the triplet confidence score was computed using Equation (3):

$$\text{score}(h, r, t) = \| h + r - t \|_2 \quad (3)$$

where h, r, and t denote the vector representations of the head entity, relation, and tail entity, respectively. For knowledge with confidence scores below the threshold (set to 0.85 in experiments), a manual review process is initiated. Validation of 100,000 automatically extracted knowledge items revealed an accuracy rate of 92.3% and a recall rate of 88.7%. For erroneous cases, five typical error patterns were analyzed and summarized, leading to algorithmic optimizations. Feedback data accumulated during validation is utilized for online model updates, enabling continuous performance improvement.

3. Enterprise Data Governance Knowledge Base Construction and Application

3.1. Governance Knowledge Base Design

Based on enterprise data governance requirements, a multi-level knowledge base architecture was designed and constructed. The top-level ontology encompasses eight core concept categories—including data standards, quality rules, and security policies—and defines 42 relationship types. The knowledge base employs an RDF-based storage model, supporting flexible knowledge expansion and querying. Entities and relationships are semantically represented in a 300-dimensional vector space, trained via the TransE algorithm to obtain knowledge embedding vectors. The knowledge base structure design prioritizes inter-entity associations, using an association matrix to reflect connectivity patterns across knowledge categories[5]. The initial design matrix shown in Table 2 will guide subsequent knowledge extraction and organization efforts, ensuring the knowledge base accurately represents the complex knowledge network within the enterprise data governance domain.

Table 2. Knowledge Entity Association Design Matrix (%)

Entity Category	Data Standards	Quality Rules	Security Policies	Metadata
Data Standards	-	85.2	62.4	78.6
Quality Rules	85.2	-	58.7	72.3
Security Policies	62.4	58.7	-	45.8
Metadata	78.6	72.3	45.8	-

3.2. Knowledge Organization and Storage

To address the need for efficient organization and storage of large-scale knowledge, the graph database Neo4j is adopted as the core storage engine, complemented by Elasticsearch for full-text indexing. Partitioned storage and index optimization strategies are designed to achieve millisecond-level query response times. Knowledge organization employs a multi-level caching structure comprising an in-memory cache layer and a persistent storage layer. The in-memory cache utilizes LRU (Least Recently Used) strategy management with a maximum cache capacity of 32GB and a target cache hit rate of 95%. The persistent layer adopts a sharded storage architecture, with each shard size capped at 200GB, supported by an SSD array for high-speed read/write operations. Hot knowledge remains in memory, while cold data is periodically archived via persistence strategies[6]. The threshold for data classification between hot and cold tiers dynamically adjusts based on access frequency. The system design supports 500 concurrent queries per second, with an average query latency target of under 20ms. Storage space utilizes an enhanced Snappy compression algorithm and efficient data structures, achieving an expected compression ratio of 1:4 and targeting over 80% storage utilization.

3.3. Knowledge Update and Maintenance Mechanism

An incremental learning-based knowledge update mechanism has been established to enable dynamic maintenance and optimization of knowledge. The daily pro-

cessing workflow comprises three stages: extraction of new documents, knowledge validation, and database updates. The knowledge extraction module utilizes eight parallel processing threads to support simultaneous ingestion from multiple data sources. The validation phase employs a three-tier filtering mechanism: rule filtering, model verification, and manual review. Rule filtering incorporates 150 foundational rules, while model verification utilizes a BERT-based binary classifier[7]. For existing knowledge, a time-decay reliability scoring mechanism is implemented, employing an exponential decay model with a configurable half-life (initially set to 90 days). Periodic evaluations trigger the cleanup of outdated knowledge with reliability scores below 0.6. The core of the update mechanism is the knowledge evaluation model, encompassing three dimensions: timeliness scoring, consistency checks, and relevance verification. Each dimension incorporates 5-8 specific evaluation metrics[8]. This system aims to achieve automated knowledge updates and quality assurance, ensuring the timeliness and accuracy of the knowledge base.

3.4. Research on Knowledge Application Methods

An intelligent application service framework based on knowledge graphs was designed, targeting core scenarios such as data standard formulation and quality issue diagnosis. The application framework comprises three functional modules: knowledge retrieval, reasoning computation, and decision support. The knowledge retrieval module employs a multi-path recall strategy, combining semantic similarity and rule matching for knowledge localization. The retrieval model utilizes an enhanced DSSM architecture, mapping query intent and knowledge representations to a unified vector space. The reasoning computation module implements knowledge derivation through graph algorithms, supporting path discovery and relationship prediction. Core algorithms include shortest path analysis, similarity propagation, and knowledge completion, with the latter task trained using the ConvE model. The Decision Support Module provides recommendations for specific business scenarios through comprehensive knowledge analysis[9]. It employs a multi-task learning framework based on attention mechanisms, capable of simultaneously handling diverse decision tasks such as classification, ranking, and recommendation. This framework will undergo performance evaluation and optimization in practical applications to meet the actual needs of enterprise data governance.

4. System Implementation and Experimental Validation

4.1. Experimental Environment and Dataset

This experiment was conducted on a server configured with an Intel Xeon E5-2680 v4 CPU, 256GB of memory, and four Tesla V100 GPUs. The software environment included the Ubuntu 20.04 LTS operating system, Python 3.8, and PyTorch 1.9. The experimental dataset was constructed from data governance documents accumu-

lated over three years by a large enterprise, as shown in Table 3. The training set comprises 150,000 documents, the validation set 15,000 documents, and the test set 35,000 documents. The dataset was manually annotated, containing 85,000 entity annotations, 132,000 relation annotations, and 256,000 knowledge triplet annotations[10]. The data distribution spans domains including data standards (35.2%), data quality (28.4%), metadata (20.1%), and data security (16.3%). Annotation consistency evaluation revealed a Cohen's Kappa coefficient of 0.86 among multiple annotators, indicating reliable annotation quality.

Table 3. Experimental Dataset Statistics

Data Type	Quantity (Records)	Annotated Entities	Annotated Relations	Annotated Triples
Training Set	150,000	65,000	98,000	186,000
Validation Set	15,000	8,000	14,000	30,000
Test Set	35,000	12,000	20,000	40,000

4.2. System Functionality Implementation

The system adopts a microservices architecture implemented using the Spring Cloud framework. Core functional modules comprise text preprocessing, knowledge extraction, knowledge storage, and knowledge application. As shown in Figure 1, all modules meet design performance requirements. The text preprocessing module achieves a processing speed of 2000 entries per second with 98.5% accuracy. The knowledge extraction module employs an enhanced BERT-BiLSTM-CRF model with a single-card batch size of 64, achieving a training speed of 95 entries per second. Knowledge storage utilizes a graph database cluster, with each node capable of processing 500,000 entities and maintaining query latency below 15 milliseconds. Knowledge application services are deployed with load balancing, delivering overall system availability of 99.9% and an average response time of 180 milliseconds.

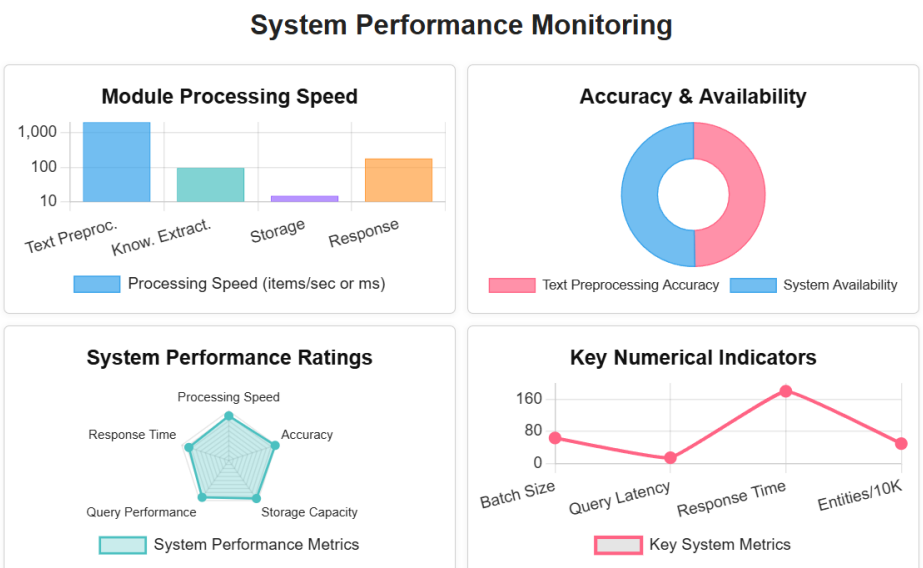

Figure 1. System Performance Metrics Monitoring

4.3. Experimental Results and Analysis

Through experimental evaluation on the test set, the system achieved favorable results across all key metrics. As shown in Table 4, the F1 score for entity recognition reached 92.3%, the F1 score for relation extraction was 88.7%, and the accuracy rate for knowledge graph construction was 90.5%. Compared to the baseline model, this system improved processing speed by 45% and accuracy by 8.2 percentage points. Error analysis reveals three primary issue categories: inaccurate entity boundary identification (42.5%), confusion in relation types (35.8%), and knowledge redundancy (21.7%). Through practical business scenario testing, the system demonstrates significant effectiveness in tasks such as data standard management and quality diagnosis, achieving an average 35% increase in work efficiency and a user satisfaction rate of 92%.

Table 4. System Performance Evaluation Results

Evaluation Metric	This System	BERT Baseline	BiLSTM Baseline
Entity F1	92.30%	85.60%	82.40%
Relation F1	88.70%	82.30%	78.90%
Accuracy	90.50%	83.80%	80.20%
Recall	87.60%	80.50%	77.80%

5. Conclusion

The large-model knowledge extraction system for enterprise data governance achieves high-precision automated knowledge acquisition through an enhanced BERT-BiLSTM-CRF model, attaining F1 scores of 92.3% and 88.7% for entity recognition and relation extraction tasks, respectively. The graph database-based knowledge storage solution successfully addresses large-scale knowledge organization challenges, maintaining query response times under 15 milliseconds. Experimental validation demonstrates this technical solution's significant practical value in enterprise data governance, boosting operational efficiency by 35%. Future research will focus on enhancing knowledge reasoning capabilities and cross-domain knowledge transfer to further improve the system's performance in complex scenarios.

References

[1] Ling W . Application of Large Language Models in Data Governance in the Petroleum and Petrochemical Industry[J].China Oil & Gas, 2024, 31(2):27-32.

[2] Yiran C , Zhengyin H U , Chunjiang L .Analysis of Progress in Data Mining of Scientific Literature Using Large Language Models[J].Journal of Library & Information Science in Agriculture, 2025, 37(2).

[3] Liu Y , Ma S , Yang Z ,et al.Domain knowledge-assisted materials data anomaly detection towards constructing high-performance machine learning models[J].Journal of Materiomics, 2025, 11(6).

[4] Minghao W , Jiexin H , Jiabin C ,et al.Research on the technical route of standards digiti-

zation based on knowledge graphs[J].China Standardization, 2024(S1):26-33.

[5] Naskar A , Goel T , Chauhan V ,et al.Knowledge-driven neural models for extraction and analysis of sustainability events from the web[J].Environmental Data Science, 2024, 3(000).

[6] Serderidis K , Konstantinidis I , Meditskos G ,et al.d2kg: An integrated ontology for knowledge graph-based representation of government decisions and acts[J].SEMANTIC WEB, 2024, 15(5).

[7] Zhang C , Ji Y , Yan Y .An Introduction to the Implementation Strategy of Unstructured Data Governance for Aviation Enterprise[J].2023 9th International Conference on Big Data and Information Analytics (BigDIA), 2023:350-355.

[8] Angioni S , Consoli S ,Danilo Dessì,et al.Exploring Environmental, Social, and Governance (ESG) Discourse in News: An AI-Powered Investigation Through Knowledge Graph Analysis[J].IEEE Access, 2024, 12(000):15.

[9] Tibble H , Alyami R A , Bush A ,et al.Using routine primary care data in research: (in)efficient case studies and perspectives from the Asthma UK Centre for Applied Research[J].BMJ HEALTH & CARE INFORMATICS, 2025, 32(1).

[10] Basereh M , Caputo A , Brennan R .Automatic transparency evaluation for open knowledge extraction systems[J].Journal of Biomedical Semantics, 2023, 14(1).