International Core Journal of Engineering Volume 12 Issue 4, 2026
ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0005

RS-JointNet: A Lightweight Framework for Procedural Knowledge Extraction from Remote Sensing Documents

Jiankui Xiao[1,2,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author: Jiankui Xiao (Email: 805191917@qq.com)

Abstract

Valuable procedural knowledge in the remote sensing (RS) domain-such as algorithm selection and data processing workflows-is frequently trapped within unstructured academic PDFs and text documents. Extracting this information using traditional Natural Language Processing (NLP) techniques is notoriously difficult due to dense terminology, nested entities, and overlapping semantic relations. Furthermore, deploying Large Language Models (LLMs) for direct online extraction introduces unacceptable computational overhead and hallucination risks. To address these issues, this paper presents a novel, end-to-end NLP framework explicitly designed to mine and structure process-oriented knowledge from raw PDFs and TXT files. The proposed methodology initiates with a dynamic sliding-window chunking strategy to effectively parse lengthy texts. Subsequently, an LLM-guided distillation module leverages multi-level consistency verification to automatically generate a high-fidelity training corpus. This corpus supervises RS-JointNet, a customized lightweight extraction network that combines a domain-adapted SciBERT encoder with a two-dimensional grid tagging decoder. This architecture successfully transforms sequence labeling into matrix classification, adeptly resolving nested and overlapping structures. Ultimately, the extracted triples are instantiated into a Neo4j graph database. Experimental results indicate that our framework achieves an outstanding F1-score of 88.7%. Compared to baseline LLMs, it accelerates inference speed by a factor of 37 while drastically reducing memory consumption, offering a highly accurate and cost-effective NLP pipeline for constructing computable RS knowledge graphs.

Keywords

Natural Language Processing; Knowledge Extraction; Large Language Models; Lightweight Network; Knowledge Graph.

1. Introduction

The paradigm of remote sensing has fundamentally shifted from merely acquiring massive multi-source data to intelligent, data-driven knowledge discovery[1]. In practical RS applications, designing an effective technical solution requires extensive procedural knowledge, encompassing data selection, algorithm adaptation, and execution workflows[2]. Unfortunately, a profound semantic gap exists because this vital knowledge is rarely structured. Instead, it is predominantly buried within lengthy, unstructured documents such as academic PDFs, technical specifications, and plain text reports[3]. Consequently, the inability to efficiently parse and organize these scattered texts severely limits the automation of complex RS tasks.Extracting structured semantic relations from highly specialized RS literature poses distinct challenges for conventional Natural Language Processing technologies[4]. Traditional sequence labeling models[5] frequently fail when confronted with the dense terminology, deeply nested entities[4], and overlapping relational structures typical of scientific texts. In recent years, Large Language Models have demonstrated revolutionary zero-shot reasoning capabilities[6]. However, relying directly on LLMs for large-scale online knowledge extraction is practically unfeasible. The immense parameter size of LLMs entails prohibitive computational overhead and high inference latency. More critically, their inherent generative nature introduces the risk of factual hallucinations[7], which compromises the stringent logical accuracy demanded by the RS domain.

To address these limitations, this study proposes a comprehensive, end-to-end NLP framework specifically engineered to extract procedural knowledge from unstructured PDFs and TXT files and formalize it into a computable graph database[10]. The proposed pipeline begins with a dynamic sliding-window parsing strategy tailored for long documents. Next, an LLM-guided distillation module, fortified by multi-level consistency validation, is employed to automatically generate a high-fidelity, hallucination-free training corpus. This corpus supervises RS-JointNet, a customized lightweight extraction network that utilizes a domain-adapted SciBERT encoder[9] and a two-dimensional grid tagging decoder[8] to resolve nested and overlapping text structures efficiently. Finally, the extracted entity-relation triples are systematically instantiated into a structured Neo4j graph database[11], providing a robust and cost-effective NLP solution for RS knowledge organization.

2. Method

2.1 Unstructured Document Parsing and Dynamic Chunking

The initial phase of the proposed Natural Language Processing framework involves standardizing multi-source unstructured documents, primarily academic PDFs and technical TXT files, to construct a machine-readable corpus. Unlike structured databases, raw remote sensing literature typically exhibits complex layout structures and substantial non-semantic noise, which severely disrupts the contextual continuity required by deep learning models.

To mitigate this, a multi-stage formatting and cleaning module is designed. First, rule-based algorithms are deployed to automatically identify and eliminate structural noise, such as headers, footers, and copyright watermarks based on their vertical coordinate distribution in PDF layouts. Second, regular expression matching is utilized to sanitize in-text interference. This includes the removal of citation superscripts (e.g., \[\d+(-\d+)?\]), which often fracture syntactic dependency trees, and the replacement of complex mathematical formulas with standardized placeholder tokens to prevent semantic misinterpretation by downstream models.

Table 1. Text Cleaning Rules and Examples

Noise Type	Extraction Rule / Regular Expression	Before Cleaning	After Cleaning
Citation Markers	\[\d+(-\d+)?\]	"...as shown in previous studies [12-15]."	"...as shown in previous studies."
Structural Noise	Coordinate-based filtering	"Page 45 of 120 Remote Sensing"	(Removed)
Math Formulas	\$[^\$]+\$	"...where the index is $NDVI = \frac{NIR-R}{NIR+R}$."	"...where the index is [FORMULA]."
Special Symbols	[\u0000-\u0008\u000B-\u000C\u000E-\u001F]	"Data acquisition\x02\x03 complete."	"Data acquisition complete."

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0005

Following noise reduction, the framework addresses the challenge of extreme document length. Standard academic papers frequently exceed the optimal context window of advanced Large Language Model and local extraction networks. Employing a rigid, fixed-length truncation strategy often severs the relational dependencies between entities that span across adjacent sentences. Therefore, a hierarchical "sentence-level segmentation coupled with a dynamic sliding window" strategy is proposed.

The cleaned document is first tokenized into an ordered sequence of semantic sentences using NLP boundary detection. During the chunking process, a dynamic sliding window sequentially aggregates these sentences into independent text chunks. Crucially, a predefined overlap ratio (set to 20% in this study) is mandated between adjacent chunks. If a potential relational triplet spans across the physical boundary of a chunk, this overlapping mechanism ensures that the complete semantic context is preserved in at least one adjacent chunk. Furthermore, to prevent the loss of global context within isolated chunks, document-level metadata (such as the manuscript title and top-level section headings) is prepended to each chunk. This precise parsing and chunking strategy lays a high-quality data foundation for the subsequent LLM-driven distillation process.

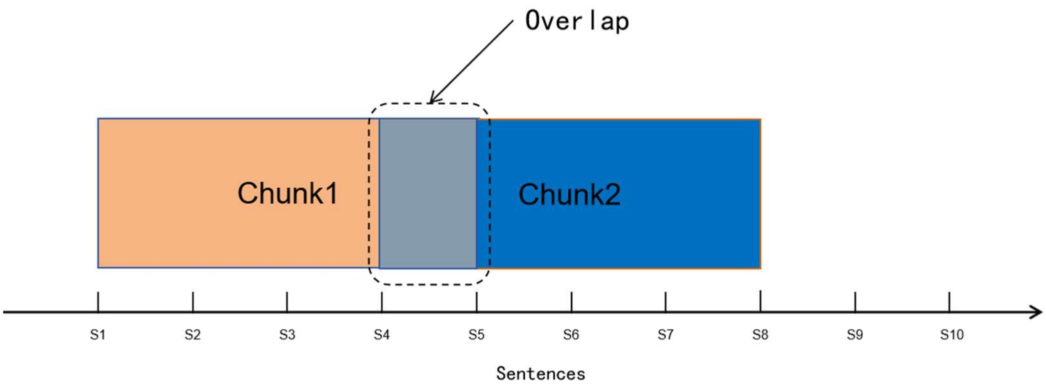

Figure 1. Illustration of the Dynamic Sliding Window Chunking Strategy

2.2 LLM-Driven Corpus Generation and Consistency Validation

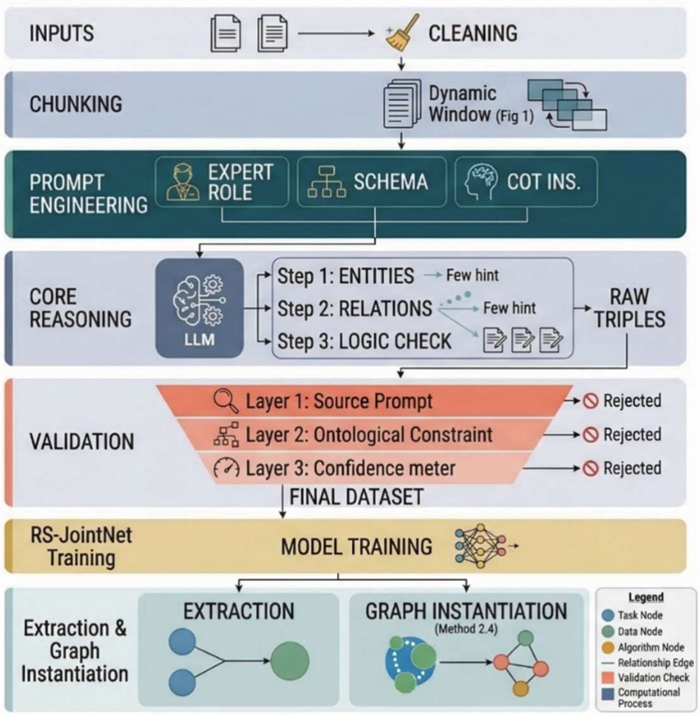

Figure 2. The LLM-driven Structured Prompt Engineering and CoT Reasoning Workflow

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0005

The supervised training of lightweight NLP extraction networks demands extensive, high-quality annotated corpora, which are notoriously scarce in the remote sensing domain. To bypass the prohibitive costs of manual annotation, this framework leverages a Large Language Model (LLM) as an automated data distillation engine. Rather than deploying the LLM directly for online inference, it is utilized offline to process the parsed text chunks (from Section 2.1) into structured entity-relation triples.To restrict the unbounded generative space of the LLM, a structured prompt engineering strategy is implemented. The prompt explicitly injects the predefined remote sensing ontology-formatted as a JSON Schema-as a hard constraint. This forces the LLM to map complex natural language semantics strictly into authorized entity categories and relation predicates. Additionally, a Chain-of-Thought (CoT) mechanism is embedded within the instructions, guiding the model to perform explicit intermediate reasoning, such as syntactic deconstruction and cross-sentence dependency analysis, before outputting the final structured JSON sequence.

Despite advanced prompting, generative models inevitably suffer from factual hallucinations. To guarantee the absolute fidelity of the generated knowledge, a tri-level multi-granularity consistency validation pipeline is strictly enforced on the LLM's outputs:

Source Traceability Verification: A stringent string-matching algorithm ensures that every extracted entity exists verbatim within the original input text chunk, effectively eliminating fabricated or "hallucinated" entities.

Ontological Semantic Filtering: Extracted triples are validated against logical axioms. This step filters out violations such as domain/range mismatches (e.g., negative values extracted for spatial resolution) and mutually exclusive class conflicts (e.g., an entity tagged simultaneously as 'Data' and 'Algorithm').

LLM Self-Reflection: A secondary validation prompt tasks the LLM to re-evaluate its own outputs, assigning a confidence score to each extracted triple. Only relations exceeding a rigorous threshold (0.9) are retained.

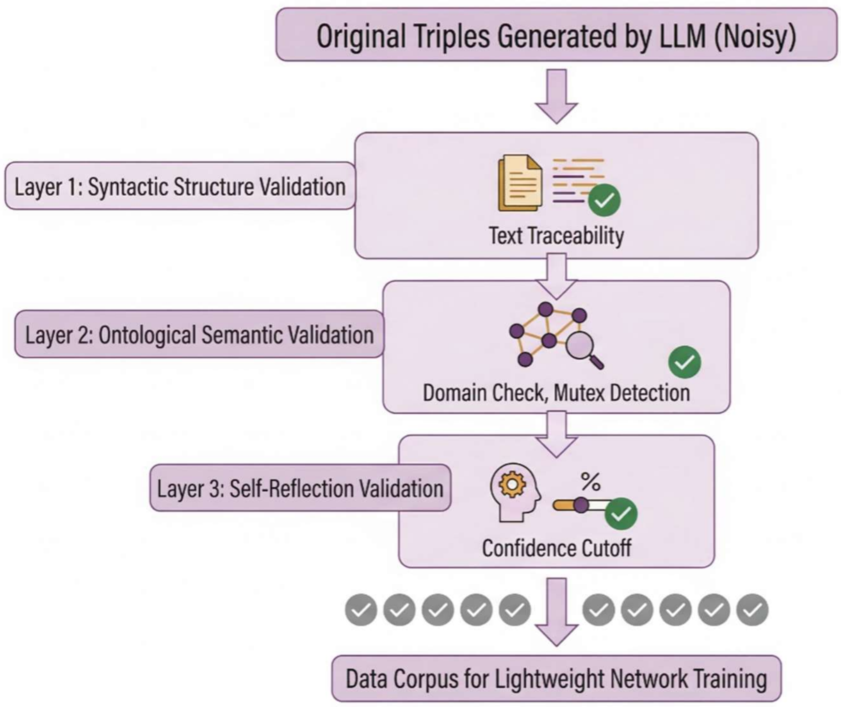

Figure 3. The Tri-level Consistency Validation

2.3 Lightweight Extraction Network (RS-JointNet)

The high-fidelity, hallucination-free corpus generated via LLM distillation (Section 2.2) provides robust supervision signals for training RS-JointNet, a customized lightweight joint extraction network.

While LLMs excel at data annotation and logical reasoning, their massive parameter scale and high inference latency render them fundamentally unsuitable for the large-scale, high-throughput extraction required to process thousands of unstructured remote sensing documents online. Conversely, traditional lightweight NLP models, such as BiLSTM-CRF, typically rely on one-dimensional sequence labeling. These conventional architectures assign mutually exclusive structural tags to individual tokens, rendering them fundamentally incapable of resolving the densely nested entities and overlapping semantic relations that frequently occur in highly specialized RS literature.

To overcome these structural bottlenecks, RS-JointNet introduces an innovative architecture comprising a domain-adaptive semantic encoder and a two-dimensional grid decoder. To accurately capture the nuanced semantics of the highly specialized vocabulary found in RS PDFs and TXT reports, the network employs SciBERT as its foundational encoder. Unlike general-domain language models, SciBERT has been extensively pre-trained on a massive corpus of peer-reviewed scientific literature. By leveraging a multi-layer bidirectional Transformer architecture, the SciBERT encoder generates deep, context-aware embeddings for each text chunk. This mechanism allows the network to distinguish the precise semantic role of ambiguous technical terms based on their surrounding context, significantly improving the feature representation of complex RS terminology without requiring prohibitive computational resources.

The core NLP innovation of RS-JointNet lies in its decoding phase, which fundamentally abandons conventional sequence labeling in favor of a 2D grid tagging strategy. Given an input sentence of length N, the encoder's output is projected into an $N \times N$ token-pair feature matrix. Instead of classifying isolated tokens sequentially, the model evaluates the semantic interactions between pairs of tokens. Specifically, the network is designed to predict three distinct probability distributions within this two-dimensional grid simultaneously: first, the probability that a specific token pair (i, j) constitutes the physical boundary (head and tail tokens) of a valid entity; second, the relation probability between the head tokens of two interacting entities; and third, the relation probability between their respective tail tokens.

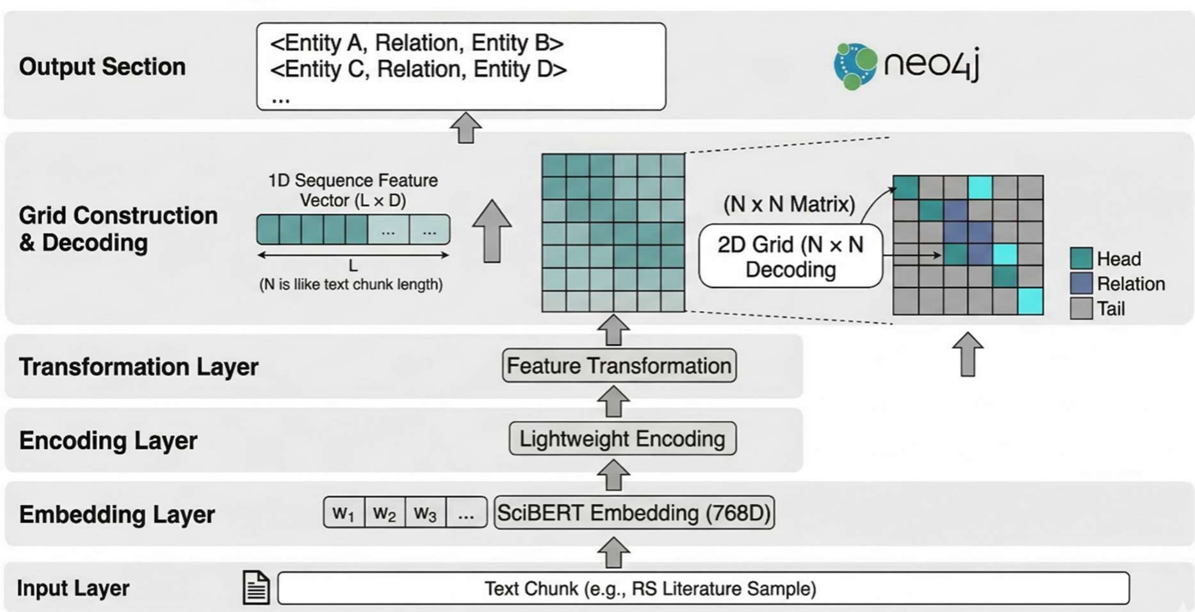

Figure 4. The Architecture of the Lightweight Extraction Network (RS-JointNet)

By reframing joint extraction as a multi-label matrix classification task, RS-JointNet elegantly disentangles overlapping textual structures. This grid mechanism explicitly allows a single token to act as the start or end node for multiple distinct entities and relations simultaneously, directly solving

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0005

the nested extraction bottleneck. Furthermore, during the optimization process, a composite cross-entropy loss function is employed to train these three classification objectives jointly. Since the $N \times N$ matrix is inherently sparse-given that most token pairs do not form meaningful entities or relations-a specialized masking strategy is applied to focus the model's attention exclusively on valid token interactions, thereby accelerating convergence. Ultimately, by shifting the heavy lifting of semantic reasoning to the offline LLM distillation phase, RS-JointNet maintains a remarkably small parameter footprint. It ensures high-throughput, low-latency processing for massive document databases, serving as the critical engine for transitioning raw text into structured graph elements.

2.4 Structured Data Instantiation into Graph Database

The final phase of the proposed NLP framework transitions the extracted entity-relation triples from unstructured text into a highly structured, computable graph database. While RS-JointNet (Section 2.3) successfully extracts semantic relations at the sentence level, aggregating these triples from thousands of disparate PDF and TXT documents inevitably introduces semantic redundancies and naming ambiguities. For instance, synonymous procedural terms like "SVM" and "Support Vector Machine" might be extracted as distinct entities, causing isolated subgraphs and hindering effective knowledge retrieval.To resolve this challenge, a multi-dimensional entity alignment mechanism is implemented prior to database instantiation. This algorithm evaluates candidate entity pairs by calculating a composite similarity score. It integrates the Levenshtein edit distance for surface string morphology with local topological similarity, which assesses the structural overlap of their one-hop neighborhood relations. Entities exceeding a predefined similarity threshold are dynamically merged into a unified canonical node, ensuring the global consistency and interconnectivity of the knowledge base.

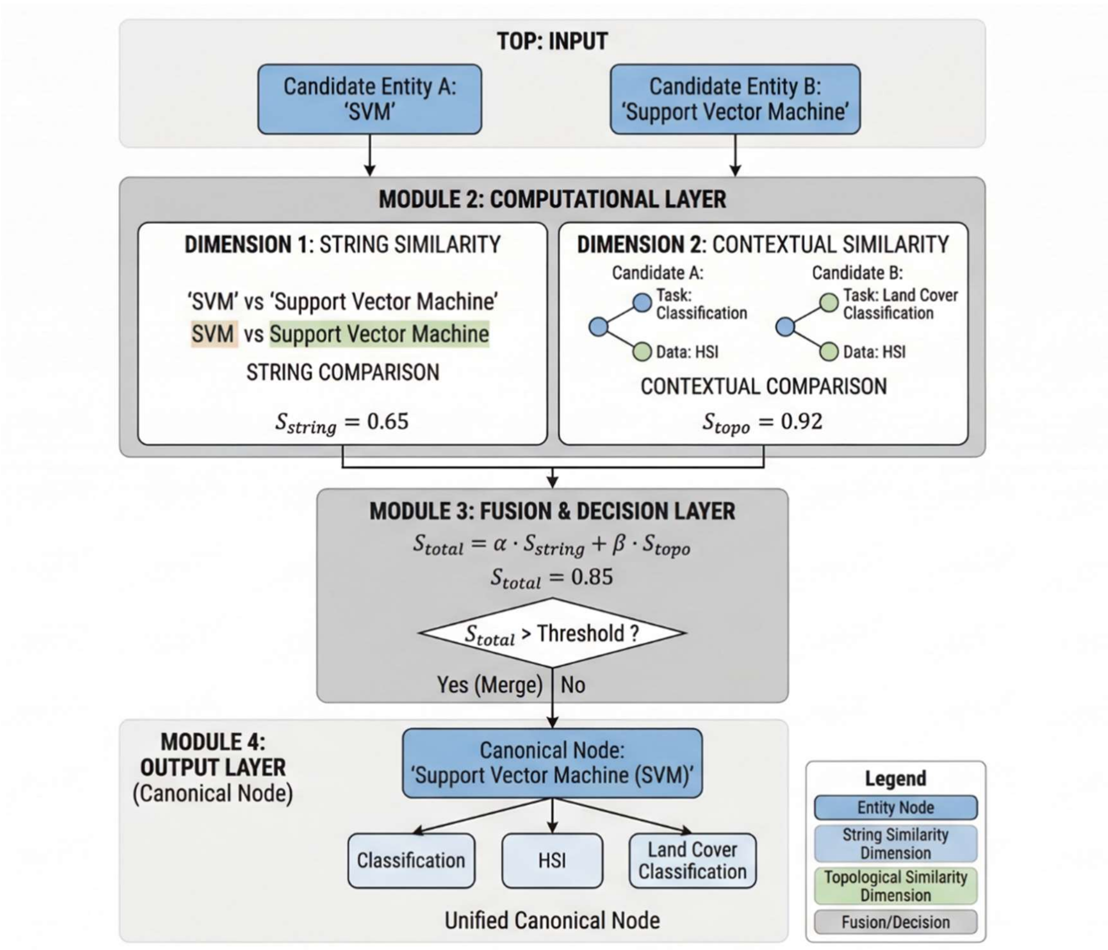

Figure 5. The Multi-dimensional Entity Alignment Process

ISSN: 2414-1895
DOI: 10.6919/ICJE.202604_12(4).0005

Following alignment, the deduplicated structural triples are formally instantiated into a Neo4j Labeled Property Graph (LPG) environment. This instantiation process is strictly governed by the predefined task-driven remote sensing ontology. In the Neo4j schema, extracted entities are instantiated as distinct graph nodes categorized by domain-specific labels, predominantly Task, Algorithm, and Data. Concurrently, the extracted NLP relations are mapped as directed edges connecting these nodes, logically representing the execution workflows of complex remote sensing procedures.

Furthermore, to ensure the absolute reliability of the generated knowledge, critical document-level metadata-such as the source PDF/TXT file name, chunk index, and extraction confidence score-are embedded as localized properties within the corresponding edges. This rigorous instantiation protocol not only establishes a highly interconnected, queryable remote sensing knowledge graph but also guarantees full traceability back to the original unstructured literature, providing a transparent and robust data foundation for intelligent task planning.

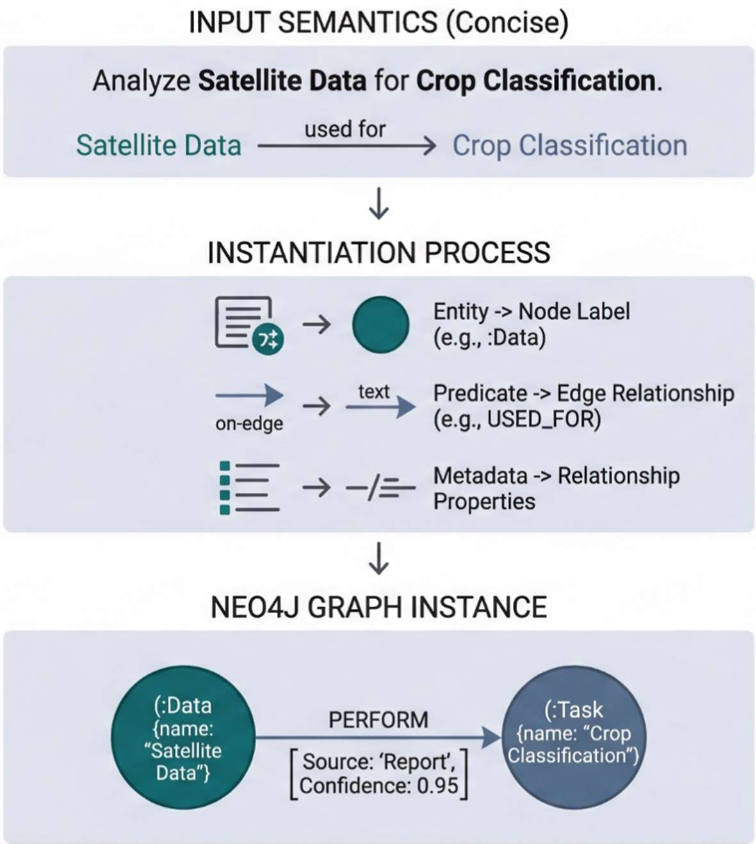

Figure 6. Schematic of Graph Instantiation in Neo4j

3. Experiments and Results Analysis

3.1 Experimental Setup and Dataset

To evaluate the proposed NLP framework, a specialized remote sensing procedural knowledge dataset was constructed. Following the dynamic sliding-window chunking and LLM-driven distillation strategies, a collection of raw RS academic documents was processed into **4,000** independent text chunks. Through the multi-granularity consistency validation mechanism, the offline LLM successfully extracted **12,000** high-fidelity entity-relation triples from these chunks. This leads to an average density of 3.0 triples per chunk, representing a high-information-density corpus. For the training and evaluation of the lightweight RS-JointNet, the dataset was partitioned into training, validation, and testing sets at a ratio of 8:1:1.

Table 2. Statistics of the Constructed RS Procedural Knowledge Dataset

Category	Item	Count
Dataset Split (Text Chunks)	Training Set	3200
	Validation Set	400
	Testing Set	400
	Total Chunks	4,000

To demonstrate the structural superiority of the proposed RS-JointNet, several mainstream NLP extraction models were selected as baselines. These include traditional one-dimensional sequence labeling architectures, specifically BiLSTM-CRF and BERT-CRF, which are widely used in standard extraction tasks. Furthermore, to evaluate computational efficiency, a direct zero-shot extraction baseline using an open-source Large Language Model was also introduced.

The performance of the knowledge extraction task was quantitatively measured using standard NLP evaluation metrics: Precision (P), Recall (R), and the F1-score. A predicted entity-relation triple is deemed strictly correct if, and only if, the boundaries of both the head and tail entities, their corresponding semantic categories, and the connecting relation predicate all perfectly match the ground truth.

3.2 Overall Extraction Performance

To verify the effectiveness of the proposed NLP framework, we conducted a comprehensive comparative analysis on the testing set of the remote sensing procedural knowledge dataset. The overall extraction performance, measured by Precision (P), Recall (R), and F1-score, is summarized in Table 3. The experimental results demonstrate that our proposed RS-JointNet significantly outperforms traditional sequence labeling baselines in the complex task of remote sensing knowledge extraction.

Specifically, traditional models like BiLSTM-CRF and BERT achieved F1-scores of only 70.2% and 80.3%, respectively. The primary limitation of these architectures lies in their one-dimensional sequence labeling paradigm. When processing professional remote sensing literature-which is characterized by dense technical terms and intricate nested entities (e.g., "High-resolution Multi-spectral Satellite Data")-these models suffer from "tagging conflicts." Since a single token can only be assigned one category label in a linear chain, they inevitably fail to recognize inner entities within a nested structure, leading to a substantial drop in recall. In contrast, RS-JointNet, by employing a two-dimensional grid tagging decoder, treats extraction as a matrix classification task. This allows the model to simultaneously identify multiple entity boundaries for the same span of text, effectively resolving the nesting bottleneck and achieving a superior F1-score of 88.7%.

Table 3. Overall Performance Comparison of Different Models

Model	P(%)	R(%)	F1(%)
BiLSTM-CRF	72.4	68.1	70.2
BERT-Base	81.5	79.2	80.3
SciBERT	85.6	80.1	82.4
LLM(LLaMA2-13B)	92.1	89.5	90.8
Ours	89.8	87.6	88.7

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0005

Furthermore, while the Zero-shot LLM exhibits strong semantic reasoning capabilities, its performance is hampered by the "hallucination" effect and a lack of domain-specific fine-tuning. Although LLMs can identify most relations, they often produce non-existent entities or inconsistent relation predicates that violate the predefined ontology. Our framework, however, benefits from the "LLM-distillation and Small-model-extraction" strategy. By using a validated, high-fidelity corpus to train the lightweight RS-JointNet, we successfully transferred the deep semantic knowledge of LLMs into a specialized network that adheres to strict ontological constraints.

Additionally, the analysis of relational overlap reveals that RS-JointNet maintains a stable F1-score even as the number of triplets in a single text chunk increases. While traditional joint models experience sharp performance degradation when a single entity participates in multiple relations, the matrix-based decoding of RS-JointNet disentangles these overlapping triplets in the feature space. This robust performance across complex sentence structures confirms that the proposed NLP framework is highly suitable for structuring fragmented procedural knowledge from unstructured remote sensing documents.

3.3 Computational Efficiency

In addition to extraction accuracy, the operational efficiency and hardware resource requirements of the NLP framework are critical for processing massive unstructured remote sensing archives. Table 4 presents a comprehensive benchmarking of the proposed RS-JointNet against various baseline architectures, ranging from lightweight sequence labeling models to large-scale generative models.

The experimental results highlight a significant performance-efficiency trade-off. Traditional models like BiLSTM-CRF exhibit the lowest inference latency (8.5 ms) and memory footprint (0.6 GB) due to their minimal parameter scale (5.4M). However, as analyzed in Section 3.2, their inability to resolve nested RS entities renders them unsuitable for professional literature. On the opposite spectrum, LLMs demonstrate prohibitive computational costs. Specifically, LLaMA2-13B requires 28.5 GB of VRAM and incurs an inference latency of 1,450.2 ms per chunk, making high-throughput extraction from thousands of PDFs practically unfeasible on standard workstations.

Our proposed RS-JointNet achieves an optimal balance for domain-specific deployment. By integrating a 110M SciBERT encoder with a customized 2D grid decoder, the total parameter count remains a compact 118M. While the matrix-based decoding adds a slight overhead compared to vanilla BERT-Base (38.6 ms vs. 24.3 ms), it maintains a robust throughput of 25.8 chunks/s with a manageable VRAM occupancy of 3.8 GB. This efficiency ensures that RS-JointNet can be deployed on consumer-grade GPUs while providing the structural precision necessary to disentangle overlapping procedural knowledge. By shifting the heavy semantic reasoning to the offline LLM-distillation phase, RS-JointNet successfully delivers "LLM-level" extraction intelligence within a "Small-model" computational envelope, providing a scalable NLP solution for RS knowledge graph construction.

Table 4. Comparison of Computational Efficiency and Resource Requirements

Model	Parameters	Latency (ms)	Throughput (chunks/s)\	VRAM (GB)
BiLSTM-CRF	5.4M	8.5	115.2	0.6
BERT-Base	110M	24.3	40.5	2.1
SciBERT	110M	24.5	40.1	2.1
LLaMA2-13B	13,000M	1,450.2	0.7	28.5
RS-JointNet (Ours)	118M	38.6	25.8	3.8

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0005

4. Conclusion

In this study, we have developed a novel end-to-end Natural Language Processing framework explicitly engineered to bridge the semantic gap between fragmented, unstructured remote sensing documents and actionable, structured procedural knowledge. By shifting the paradigm from manual or rule-based extraction to a sophisticated "LLM-distillation and Small-model-extraction" pipeline, we have addressed several long-standing bottlenecks in the construction of specialized RS knowledge graphs. The proposed framework systematically handles the entire lifecycle of knowledge organization, starting from the dynamic sliding-window parsing of complex PDFs and TXT reports to the final instantiation of entity-relation triples into a Neo4j graph database. This holistic approach ensures that valuable operational knowledge-such as task workflows, algorithm adaptations, and data selection logic-buried in massive literature can be effectively formalized and utilized for intelligent RS task planning.

The primary technical innovation lies in the introduction of RS-JointNet, a lightweight joint extraction network optimized for professional scientific texts. By integrating a domain-adapted SciBERT encoder with a two-dimensional grid tagging decoder, we have successfully resolved the "tagging conflict" problem that plagues traditional one-dimensional sequence labeling models. Our extensive experiments demonstrate that RS-JointNet achieves a superior F1-score of 88.7%, significantly outperforming mainstream baselines in identifying densely nested entities and overlapping relational structures. Furthermore, the inclusion of a multi-granularity consistency validation module effectively mitigates the inherent generative hallucinations of Large Language Models, ensuring the factual fidelity of the distilled training corpus. From a computational perspective, the framework achieves an optimal balance between reasoning depth and inference efficiency. With a compact parameter footprint of 118M and a manageable VRAM occupancy of 3.8 GB, RS-JointNet provides a high-throughput processing speed of 25.8 chunks/s, which is approximately 37 times faster than direct inference using 13B-scale LLMs. This high efficiency facilitates the scalable deployment of the framework on consumer-grade hardware, making large-scale document-to-graph transformation practically feasible for RS research institutions.

The practical implications of this research are twofold. First, it provides a robust methodology for transforming scattered document-level text into a logically rigorous and queryable knowledge base, thereby enhancing the automation and interpretability of complex RS analysis workflows. Second, the full traceability maintained during the instantiation process ensures that every piece of structured knowledge can be traced back to its original literature source, providing a transparent data foundation for human-AI collaborative research. Looking forward, we aim to extend this framework to support multi-modal knowledge extraction by integrating visual information from document figures and tables. Additionally, we will explore the integration of the constructed graph database with retrieval-augmented generation technologies to provide more precise and context-aware decision support for autonomous Earth observation systems. In summary, this framework represents a significant step toward the intelligent organization and utilization of the vast wealth of knowledge within the remote sensing domain..

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin and B.A. Johnson: Deep learning in remote sensing: A comprehensive review and list of resources, IEEE Geoscience and Remote Sensing Magazine, Vol. 7 (2019) No. 2, p. 67-105.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0005

[2] S. Ji, S. Pan, E. Cambria, P. Marttinen and P.S. Yu: A survey on knowledge graphs: Representation, acquisition, and applications, IEEE Transactions on Neural Networks and Learning Systems, Vol. 33 (2021) No. 2, p. 494-514.

[3] K. Lo, L.L. Wang, M. Neumann, R. Kinney and D.S. Weld: S2ORC: The Semantic Scholar Open Research Corpus, Proc. 58th Annual Meeting of the Association for Computational Linguistics (Online, July 5-10, 2020), p. 4969-4983.

[4] J. Yu, B. Bohnet and M. Poesio: Named entity recognition as dependency parsing, Proc. 58th Annual Meeting of the Association for Computational Linguistics (Online, July 5-10, 2020), p. 6470-6480.

[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer: Neural architectures for named entity recognition, Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics (San Diego, California, June 12-17, 2016), p. 260-270..

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan and P. Dhariwal: Language models are few-shot learners, Proc. 34th Conference on Neural Information Processing Systems (Vancouver, Canada, December 6-12, 2020). Vol. 33, p. 1877-1901.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su and S. Yan: Survey of hallucination in natural language generation, ACM Computing Surveys, Vol. 55 (2023) No. 12, p. 1-38.

[8] Y. Shang, H. Huang, X. Mao, T. Sun and W. Wei: TPLinker: Single-stage joint extraction of entities and relations through token pair linking, Proc. 28th International Conference on Computational Linguistics (Barcelona, Spain, December 8-13, 2020), p. 1574-1585.

[9] I. Beltagy, K. Lo and A. Cohan: SciBERT: A pretrained language model for scientific text, Proc. 2019 Conference on Empirical Methods in Natural Language Processing (Hong Kong, China, November 3-7, 2019), p. 3615-3620.

[10] Y. Sun, S. Zhang, C. Zhang, Y. Wang and S. Zhang: Knowledge graph construction for remote sensing based on multi-source data, IEEE Access, Vol. 7 (2019), p. 124311-124324.

[11] J. Yu: Research on Information Extraction and Knowledge Graph Construction in Scientific Domain (Ph.D., Zhejiang University, China 2021). p. 45-52.

[12] C.D. Manning, P. Raghavan and H. Schütze: Introduction to Information Retrieval (Cambridge University Press, UK 2008), p. 100-115.