ISSN 1392-3730/eISSN 1822-3605

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of CIVIL ENGINEERING and MANAGEMENT

2026
Volume 32
Issue 4
Pages 548–562
https://doi.org/10.3846/jcem.2026.26519

ONTOLOGY DEVELOPMENT USING LANGUAGE MODEL-BASED NAMED ENTITY RECOGNITION FOR INTEGRATED CONSTRUCTION INFORMATION

Goeun CHOI [1], Soonwook KWON [2]✉, Jinwoo SONG [3], Ali AKBAR [1], Jung-taek HONG [1]

[1] *Department of Global Smart City, Sungkyunkwan University, Gyeonggi-do, South Korea*
[2] *School of Civil, Architectural Engineering and Landscape Architecture, Sungkyunkwan University, Gyeonggi-do, South Korea*
[3] *Department of Civil and Environmental Engineering and Institute of Construction and Environmental Engineering, Seoul National University, Seoul 08826, South Korea*

Article History:
- received 28 March 2025
- accepted 8 December 2025

Abstract. Named Entity Recognition (NER) is crucial for building knowledge bases and facilitating semantic search in the construction industry. While conventional NER models can identify general entities such as spatial and organizational information, extracting domain-specific entities, like materials and dimensions from construction-related texts –particularly in Bill of Quantities (BoQ) and Building Information Modeling (BIM) parameters – remains challenging extensive manual annotation.

Key entity categories were defined, and datasets from four BoQ and two BIM sources were annotated to establish ground truth labels. A semi-automated labelling process was introduced to streamline annotation and improve training efficiency. Experimental results demonstrate that the proposed framework reduces annotation time by nearly threefold compared to manual processes. This study developed a BERT-based NER model achieving F1 scores ranging from 0.81 to 0.97, with higher performance for well-defined construction parameters (name, material, size, thickness, diameter, length, type: 0.95–0.97) compared to miscellaneous text entities (0.81).

Despite extensive research in construction NLP, existing approaches fail to address the integration challenges between heterogeneous BIM-BoQ data formats and lack domain-specific entity recognition capabilities. The extracted entities are aligned with standardized formats using semantic text similarity techniques. This ontology-based integration enhances data consistency, interoperability, and retrieval accuracy, improving semantic alignment while minimizing discrepancies from heterogeneous terminology.

Keywords: Large Language Models (LLM), Natural Language Processing (NLP), Named Entity Recognition (NER), deep learning, Building Information Modeling (BIM), construction data integration, ontology development, data standardization.

✉Corresponding author. E-mail: *swkwon@skku.edu*

1. Introduction

Digital Twin technology has emerged as a key innovation in the construction industry, playing a significant role across the design, construction, and maintenance phases of the project lifecycle. However, compared to other industries, the implementation of Digital Twin in the construction sector faces several challenges. The primary obstacles can be categorized into three main issues. 1) Construction projects generate vast amounts of data from diverse sources, 2) the lack of standardized data formats leads to poor interoperability between heterogeneous systems, and 3) the insufficient integration of various types of construction-related information (Halmetoja, 2022).

The lack of data standardization and integration is a key issue hindering the development of digital twins in the construction industry. Analysis of the national digital twin pilot project identified inadequate standardization as a key issue, which is linked to service model scalability (Jeong et al., 2024). Currently, most digital twin implementations remain at the level of simple 3D modeling visualization or IoT monitoring, necessitating standardization across each industry (Yun & Kim, 2022). For the combined utilization of individual digital twins created across different domains, metadata standards are essential to ensure interoperability (Na & Kim, 2024). In the public sector, link-

ing and expanding digital twins across national territories is also identified as a key improvement priority, demonstrating the urgent need for data standardization and enhanced connectivity (Kim et al., 2020).

These issues reduce the applicability of digital twins, and to address this problem, Natural Language Processing (NLP) technology becomes a critical issue, especially for processing text-based information (specifications, reports, statements, etc.). Although the construction industry possesses vast amounts of textual data, existing technologies for efficient data processing remain inadequate (Jagannathan et al., 2022). Therefore, it is essential to develop methodologies that can extract and standardize meaningful information from unstructured text data (Wu et al., 2022b).

This study proposes the application of Named Entity Recognition (NER) to extract specific entity information from large-scale text data in the construction industry. NER has been widely validated in the field of NLP as an effective information extraction method. When applied to construction, NER offers several advantages. It enables the structuring of unstructured text data by automatically identifying entities such as locations, materials, workforce, and processes in construction project documents, thereby enhancing data usability (Jeon et al., 2022).

Furthermore, it facilitates data standardization by integrating identical concepts across various document formats (Wu et al., 2022a). Additionally, by improving the data quality of Digital Twin models, it enables more precise information integration and analysis, thereby enhancing the reliability of decision-making processes (Zhang et al., 2023).

Despite these advancements, challenges remain in applying NER to construction document text. For instance, texts in Bills of Quantity (BoQ) descriptions and BIM parameters often contain construction-specific abbreviations, inconsistent terminology, and contextual ambiguities, leading to interpretation challenges (Kuiper & Duffield, 2018). Addressing these issues requires customized solutions, including the development of high-quality ground truth labels, optimization of entity recognition models for specialized contexts, and integration of ontology-based data. Furthermore, the importance of an ontology-driven approach for comprehensive project data management has been emphasized. The interoperability between BIM systems and other documents has emerged as a critical research area, highlighting the need for standardized data exchange formats and structured contextual data systems.

The objective of this study is to extract text-based parameter information from BIM and BoQ documents, standardize the extracted data, and establish an ontology-based integrated metadata framework. This research aims to structure unstructured data in the construction industry and enhance data interoperability. The key contributions of this study are as follows. First, a construction domain-specific NER model is developed to automatically extract parameter information from BIM and BoQ documents, generating structured entity data. Second, a mapping methodology using Semantic Text Similarity (STS), which quantifies the degree of semantic equivalence between textual entities, is proposed to align extracted entities with standard data formats, thereby improving data consistency and usability. Finally, ontology is designed to integrate various data sources, enabling efficient information retrieval and utilization within the construction industry.

2. Literature review

2.1. NER development process

Named Entity Recognition (NER) has established itself as a fundamental task in Natural Language Processing (NLP), playing a crucial role in extracting valuable information from large-scale textual datasets. By identifying and categorizing entities such as names, locations, dates, and domain-specific terms, NER facilitates a wide range of applications, including information retrieval, semantic analysis, and knowledge graph construction (Pakhale, 2023).

The evolution of NER has been characterized by a transition from rule-based systems, which heavily relied on predefined patterns and linguistic rules, to more sophisticated machine learning and deep learning techniques. Early methods, such as Long Short-Term Memory (LSTM), Bidirectional-LSTM (Bi-LSTM), and Conditional Random Fields (CRF) with static embeddings, were widely adopted and demonstrated reasonable performance. However, these approaches were often constrained by their dependence on handcrafted features, domain-specific challenges, and limitations in handling low-resource languages with insufficient training data (Sammet & Krestel, 2023; Taher et al., 2020; Zhang et al., 2023).

In contrast, recent years have witnessed groundbreaking advancements in this field through pre-trained language models based on Transformer architectures, particularly BERT. These models leverage contextual embeddings to capture the semantic and syntactic nuances of text, significantly enhancing NER performance. Traditional NER models employing static embedding vectors, such as Word2Vec and Bi-LSTM architecture, have offered lightweight models with fast training and inference (Lê et al., 2019; Yang & Xu, 2020). However, they struggled to incorporate contextual information, making it difficult to resolve polysemy, infer the meaning of unseen words, and capture long-term dependencies in lengthy sentences (Luo et al., 2019). In contrast, BERT effectively addresses these limitations by utilizing dynamic embeddings that consider inter-sentence context and integrating long-range dependencies, making it particularly suited for handling complex NER tasks involving nested and discontinuous entities (Cho & Lee, 2019; Luoma & Pyysalo, 2020; Xie, 2024). In clinical NER tasks, encoder-based models such as BERT achieved superior F1-scores of 0.87–0.88 compared to LLMs' 0.18–0.30, with decoder-based LLMs showing poor recall despite high precision (Arzideh et al., 2025). BERT

demonstrates advantages in logical reasoning, being immune to the "reversal curse" that affects decoder models like GPT and performs better on complex logical reasoning tasks (Wu et al., 2024).

Contextual embeddings have proven superior to traditional word embeddings by adapting to the context of individual instances. Studies have reported up to a 13% improvement in micro F1 scores for such tasks using BERT (Lester et al., 2020; Taillé et al., 2019). Moreover, domain-specific adaptation of BERT has yielded exceptional results in fields like biomedical, financial, and legal NER, where domain-specific semantics and regulations pose significant challenges (Keshavarz et al., 2022; Pakhale, 2023; Y. Zhang & H. Zhang, 2023).

2.2. NLP and data integration in construction

With the increasing adoption of digital technologies such as BIM, the Internet of Things (IoT), and Building Automation Systems (BAS), vast amounts of data are being generated. However, data silos within the Architecture, Engineering, and Construction (AEC) industry hinder the exchange of information between buildings and innovative applications due to the lack of a common data representation (Tang et al., 2022). To address this issue, research in NLP has focused on structuring, systematizing, and integrating text data from sources such as BIM and construction specifications using NER and hierarchical structures.

NLP and Named Entity Recognition (NER) techniques play a crucial role in extracting user-intended entity data from large-scale construction documents. Extensive NER research has been conducted to extract relevant information from construction-related texts. For example, an automated compliance checking (ACC) system for BIM was proposed by extracting and standardizing data via Dynamo and utilizing BERT embeddings with a BiLSTM-CRF model to identify regulatory entities from construction specifications, enabling rule-to-model data matching (Li et al., 2024). Similarly, a hybrid deep learning model was introduced to automate constraint modeling for Advanced Work Packaging (AWP) by extracting constraint-related entities from specifications using a BiLSTM-CRF model and structuring inter-constraint relationships through a knowledge representation learning (KRL) model to generate AWP graphs (Wu et al., 2021). Moreover, a BiLSTM-CRF model was applied to extract entities, including organizations, equipment, regulations, and technical terms, from Chinese construction documents (Zhang et al., 2023).

To construct hierarchical data structures, various ontology-based or data-schema approaches have been proposed, including BIM data structuring using Industry Foundation Classes (IFC) standards enabling efficient data retrieval from BIM object databases through a natural language-based search engine (Wu et al., 2019). Similarly, a methodology was developed to automatically extract BIM project-specific properties and integrate them into an ontology, enhancing data management and reusability while employing a Sentence-BERT (SBERT)-based approach for synonym retrieval (Yin et al., 2024). Additionally, a Text-to-BIMQL framework was introduced, which utilizes Word-2Vec embeddings and graph neural networks (GNN) to convert natural language queries (NLQ) into BIM-specific structured queries, facilitating efficient data retrieval (Yin et al., 2023). Ontology applications were further expanded by developing a formalized spatial representation for transport assets and utilities, analyzing hierarchical semantic structures, and examining logical relationships in textual descriptions to improve information extraction strategies (Xu & Cai, 2021).

Despite these advancements, existing studies present certain limitations when considering comprehensive data integration for digital twin implementation. Single-system information processing studies have demonstrated excellent performance within individual domains or single document types, yet their scope remains focused on processing within specific systems rather than achieving inter-system data integration. Document comparison and classification studies have achieved advanced approaches and high performance, but are primarily designed for homogeneous data formats, which limits their applicability to resolving structural disparities between heterogeneous data sources.

Furthermore, graph-based approaches have shown promise for representing construction knowledge. The ifcOWL ontology has been developed to represent IFC schemas in OWL format, enabling semantic queries across building data (Beetz et al., 2009). RDF (Resource Description Framework) and SPARQL queries have been applied to building information, primarily focusing on structured BIM data rather than unstructured text from documents (Rasmussen et al., 2021). However, these implementations typically rely on manual ontology construction using tools like Protégé, requiring extensive domain expertise and time investment.

This approach automates entity extraction via NER while preserving semantic relationships through STS-based alignment, enabling automatic mapping between unstructured construction documents and structured ontological representations – a gap that tools like Protégé cannot automatically bridge.

This study extends existing research to extract parameter-related entities from construction documents while addressing inter-system connectivity challenges, proposing an BERT-NER-based ontology and STS integration methodology that enables seamless data connectivity for digital twin implementation. The extracted entities will be integrated into an ontology using Structured Text Similarity (STS) based on Cost Breakdown Structure (CBS) and Engineering Breakdown Structure (EBS), facilitating comprehensive data integration necessary for digital twin construction through structural integration between heterogeneous systems such as BIM models and BoQ documents.

3. Methodology

Ontology is defined as a formal and explicit specification of a shared conceptualization, enabling the representation of domain-specific knowledge in a machine-readable and processable format while incorporating defined constraints (Gruber, 1993). Additionally, ontology models semantic hierarchies by structuring and formalizing properties and relationships within classes and instances.

This study aims to construct an ontology for elements that constitute BoQ and BIM data and to automatically extract ontology-defined entities using BERT-based NER. Ontology not only provides a structured approach to representing conventions but also enables dynamic integration and analysis of entities extracted from BIM and BoQ with Cost Breakdown Structure (CBS) and Engineering Breakdown Structure (EBS) datasets. By hierarchically organizing entities and establishing clear relationships, ontology facilitates efficient classification and linkage between BIM parameter data and BoQ description data. Furthermore, textual data processing extends beyond simple pattern matching, enabling conceptual associations between integrated datasets and the analysis of semantic relationships within contextual information.

The research framework of this study is illustrated in Figure 1 and follows the following process:

(1) To integrate BIM and BoQ information, CBS and EBS classes and properties are defined, establishing the foundation for ontology-integrated data.

(2) BoQ datasets from four projects and BIM textual data from two projects are used to create training and testing datasets for NER (Table 1), following the predefined standards in step (1). The BoQ and BIM datasets primarily focus on foundational construction, steel-frame construction, and reinforced concrete (RC) construction.

(3) Through this process, entity data constituting CBS and EBS within BoQ and BIM can be identified.

(4) Finally, data acquired in step (3) is cross-referenced with CBS defined in step (1) for ontology construction and integration, using Semantic Text Similarity (STS) – a computational method that measures the semantic equivalence between text segments to establish relationships across different data sources.

3.1. Metadata for ontology

This section defines the hierarchical structure of metadata constituting the ontology, as well as CBS, the standardized text that forms this hierarchy. CBS can be structured according to various criteria based on the needs of project practitioners and organizations. In this study, one BoQ from the four analyzed projects was selected, and the types of specification information were first classified. BoQ text was categorized into seven classes: name, material, size, type, thickness, diameter, and others. Subsequently, text corresponding to specification information and stop

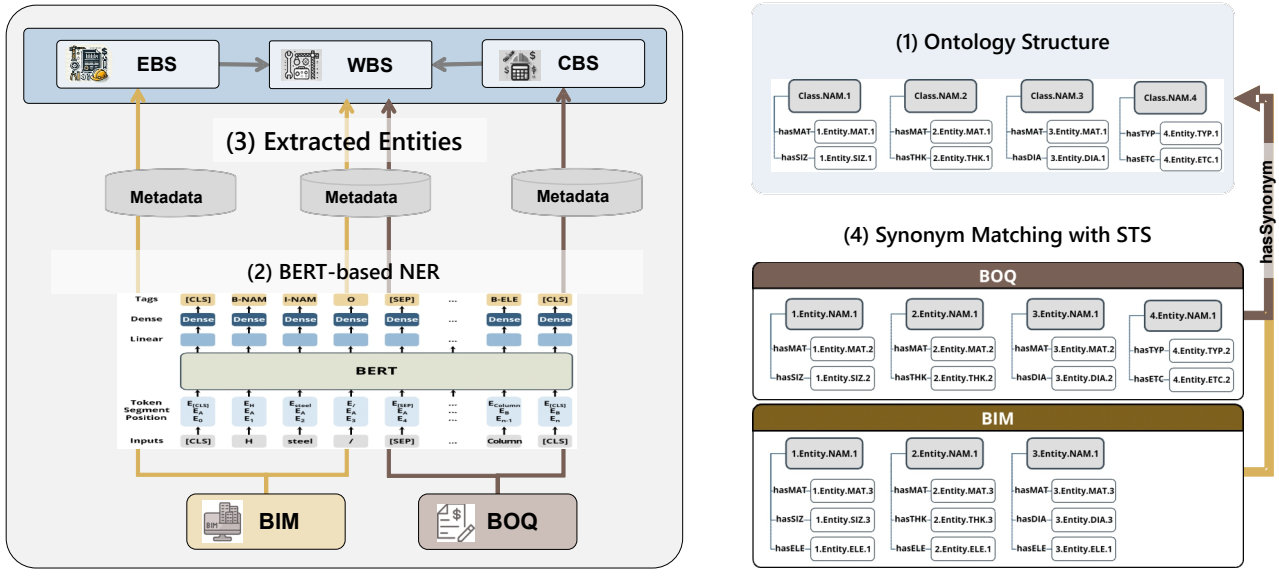

Figure 1. Research framework for ontology development using BERT-based Named Entity Recognition in construction information integration

Table 1. Project information

Project	Country	Language	Building Type	BoQ data	BIM data
PJT1	South Korea	Korean, English	Battery Manufacturing Facility	1,160	0
PJT2	South Korea	Korean, English	Battery Office Complex	122	8,474
PJT3	Indonesia	English	Battery Manufacturing Facility	51	20,220
PJ4	Poland	English	Battery Manufacturing Facility	92	0
Total				1,425	28,694

words (e.g., special characters) were removed from item names to define CBS, which consists of standardized item names and a specification template.

The selection criteria for BoQ were based on excluding three projects written in foreign languages. Among the remaining two projects, the one with greater scale and functional significance was chosen, incorporating feedback from relevant project stakeholders. The development of CBS represents a critical preliminary step in this study, as the definition of CBS item names directly affects the annotation of specification labels. For example, if the BoQ description "Auger Crane Equipment Cost, 200 ton" is defined under CBS as "Auger Crane Equipment Cost", the parameter information should be annotated as "Other: 200 ton". However, if the CBS item name is defined as "Equipment Cost", the parameter information should be annotated as "Type: Auger Crane" and "Other: 200 ton".

The hierarchical structure in this study aims to establish a structured relationship by linking EBS (Element Breakdown Structure) to CBS (Cost Breakdown Structure). EBS consists of element parameters in BIM, such as columns, beams, and walls, and is further subdivided into smaller units than CBS items, which include construction tasks like steel member installation and concrete pouring. Additionally, BoQ items that are not explicitly modeled in BIM result in a broader range of CBS items. Analyzing BIM parameters data revealed six key parameters – material, size, type, thickness, diameter, and element – which were subsequently defined as EBS components. While EBS shares common elements with CBS, such as material, size, type, thickness, and diameter, the "element" parameter is unique to EBS, containing additional information distinct from CBS. A visualization of the CBS-EBS relationship is illustrated in Figure 2.

The comprehensive parameter dataset, encompassing the primary parameters of CBS and EBS, serves as an intermediate layer standard for aligning heterogeneous BIM and BoQ datasets. This allows multiple BoQ items and BIM elements to be represented as a single CBS and EBS category through many-to-one mapping, providing the possibility of cross-system validation of BIM quantities and BoQ quantities. The names and parameter information defined for CBS and EBS in this study are extracted as ontology classes and properties. The extracted classes and properties are presented in Table 2.

3.2. NER model development process

The NER model in this study is developed following the process outlined in Figure 3. First, based on the CBS defined in the previous section, a manual annotation process is conducted to create the initial training and test datasets. The annotation follows the guidelines specified in "Metadata for Ontology". Using this dataset, the first NER model is trained, after which an entity prediction algorithm is applied to generate temporary labels automatically for the cleaned data of three BoQ projects and two BIM projects.

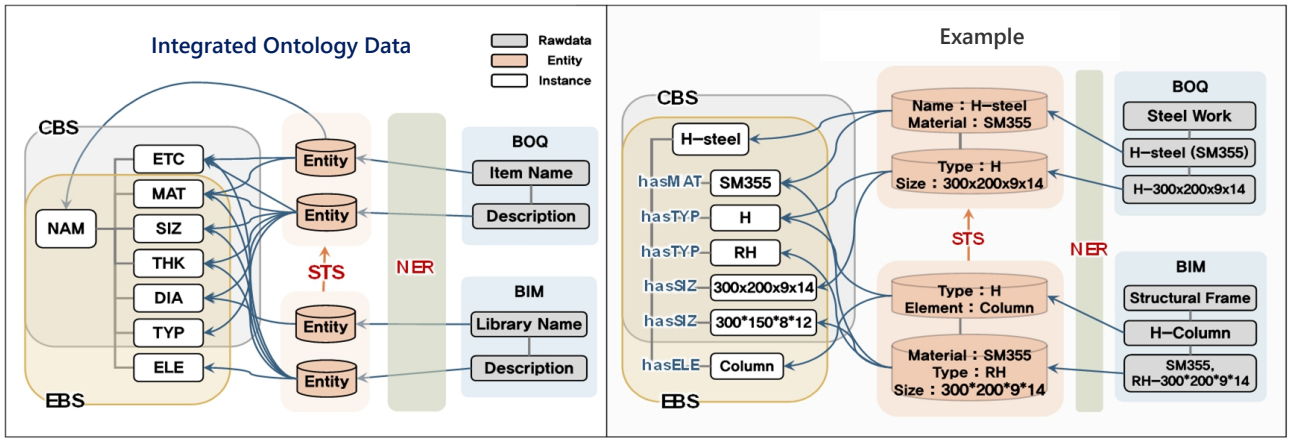

Figure 2. Example of BoQ and BIM data integration through ontology-based entity mapping. The framework shows how CBS and EBS entities are extracted via NER and integrated using STS, demonstrated through H-steel structural elements from both data sources

Table 2. Ontology class categorization for entity extraction

Class	Instance	Example	Property	Description
NAM	Name	Pile, H-steel	hasSYN	CBS/EBS has a synonym named NAM
ELE	Element	Foundation, Column, Beam	hasELE	EBS has element entity named ELE
MAT	Material	Concrete, SM355	hasMAT	CBS/EBS has material entity named MAT
SIZ	Size	300x200x9x14	hasSIZ	CBS/EBS has size entity named SIZ
THK	Thickness	6t	hasTHK	CBS/EBS has a thickness entity named THK
DIA	Diameter	D600	hasDIA	CBS/EBS has diameter entity named DIA
TYP	Type	Micro-Pile, PHC-Pile, H, L	hasTYP	CBS/EBS has type entity named TYP
ETC	Etc.	(Random Text)	hasETC	CBS has etc. entity named ETC

The data cleaning process involves handling imbalanced data, translation, and special character processing, with further details provided in the "Data Pre-processing" section. The automatically generated temporary labels were evaluated using the first model, achieving an F1 score of approximately 83%. The decrease in score can be attributed to differences in text composition, specifically: (1) BoQ and BIM data contain stopwords, while CBS is a pre-processed dataset with stopwords already removed, and (2) BIM data includes element names and library names that are absent in CBS. These discrepancies lead to inconsistencies in NER model performance when analyzing BoQ and BIM text based on CBS-trained models. After manually correcting the wrong temporary labels which rate 17% of whole data, the dataset is retrained along with the initial training and test data to develop the second NER model.

While a total of 3,610 sentences require annotation for the manual annotation process, the semi-automated approach demonstrates a substantial improvement in efficiency, reducing the amount of manually labeled data by approximately 83% in this study. Since only the remaining 17% of data with incorrect temporary labels from the first model requires manual correction, while the accurately predicted 83% can be directly utilized for training the second model.

3.3. NER model architecture

Bidirectional Encoder Representations from Transformers (BERT) is a neural network model designed for NLP tasks that captures word relationships bidirectionally. In this study, a pre-trained BERT model with 110 million parameters is fine-tuned for the specific task of entity recognition to extract contextual features and enhance NER performance. The NER model architecture, which takes a corpus as input and produces tagged output, is illustrated in Figure 4. The process of loading the pre-trained BERT model is defined as follows: H = BERT(X) where H represents the final hidden state of BERT, and X denotes the input token embedding vectors. The final hidden state H is subsequently processed using a linear transformation and a softmax function to derive the probability distribution for entity classification.

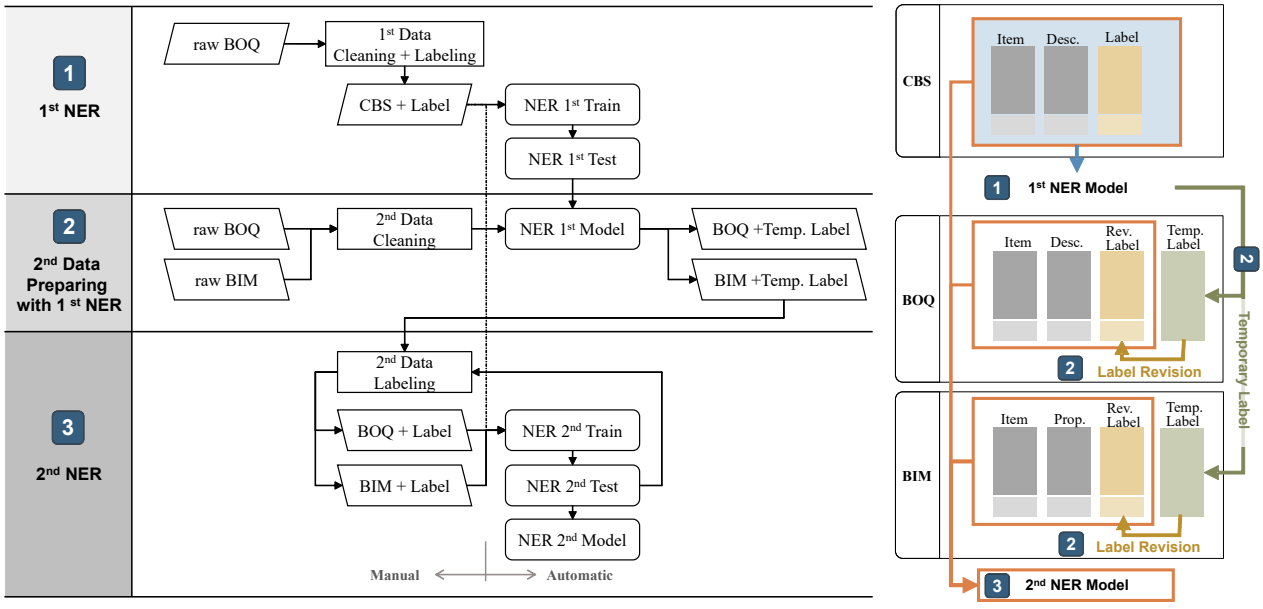

Figure 3. NER model development process using semi-automated labeling. Three-stage workflow: (1) initial model training with CBS data; (2) automated label generation with initial model; (3) final model training with BoQ and BIM datasets

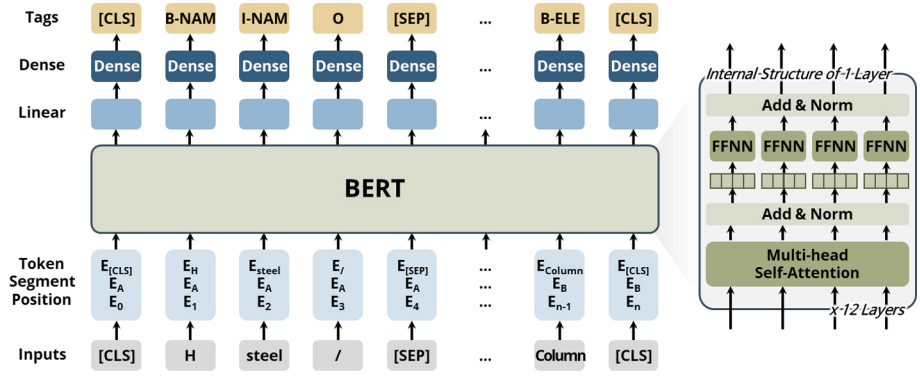

Figure 4. BERT-based NER Model Architecture. Input tokens are processed through BERT transformer layers to generate contextualized embeddings, followed by dense classification layers that output BIO-tagged entity predictions for domain-specific construction terms

3.4. Data pre-processing

To ensure effective entity extraction from construction documents, a structured data pre-processing pipeline was implemented. The data pre-processing consists of two primary stages: cleaning and labeling.

For labeling, entities and relationships were manually annotated for BoQ descriptions based on the categories defined in Section 3.1. The initial ground truth dataset served as the training set for the first BERT-based NER model, which was then used to predict entities for the remaining three BoQ descriptions and two BIM parameter datasets. The labeled data was then refined through a semi-automated correction process, and a second NER model was developed.

In the data cleaning process, given the unique characteristics of construction documents, we addressed imbalanced data, ensured language consistency, and applied rules for special characters. Notably, stopword removal and case normalization were not applied. Stopwords were labeled as "O" to retain their presence in the dataset without affecting model learning, and an uncased pre-trained BERT model was used to eliminate the need for case normalization.

1. Handling Imbalanced Data: Due to the structured nature of BIM, parameters such as component properties frequently appear in large quantities, resulting in highly redundant text. Unlike BoQ, where item quantities are represented numerically per item, BIM software characteristics cause text to be repeated as many times as the number of modelled components. To mitigate this imbalance, BIM parameters with more than four repetitions were removed, while BoQ text was augmented threefold using a decoder-based transformer model.
2. Special Character and Whitespace Handling: A consistent rule set was applied to process special characters and whitespace. Special characters were either modified during cleaning or used as tokenization markers in the NER model. Initial tokenization was conducted using whitespace and commas, forming the basis for secondary tokenization via the BERT tokenizer. Since labelled entities were extracted based on the first tokenization stage, this process was crucial. Tokenization rules were applied to process the text effectively. Whitespace and commas served as the primary tokenization markers. Special characters were manually reviewed in Excel to determine whether they required separation. Parentheses were replaced with commas to improve parsing. A forward slash was inserted between names and descriptions to enhance NER accuracy for names. For name text, a hyphen was used to join necessary words, while for description text, an underscore was used instead of spaces to maintain word integrity. For example, the text "H-steel/SM355, H-300×200×9×14" was processed following these rules to minimize manual data modification. The number of NER training & test data is shown in Table 3 for each project.
3. Language and Terminology Standardization: Construction terminology exhibits significant lexical variation, with the same concept expressed through multiple languages, synonyms, and abbreviations. This variability degrades NER model accuracy. Therefore, a two-step standardization process was applied: terminology standardization via a synonym dictionary followed by language standardization through a translation API. For instance, "Reinforced Concrete", "R.C", "철콘"(RC in Korean), and "철근콘크리트" (Reinforced Concrete in Korean) were all unified to "RC". To address this, a synonym dictionary was developed to map variant terms to standardized forms. During preprocessing, variant terms were compiled in two stages. Korean variant terms were identified by comparing pre- and post-translation outputs, while English variant terms were compiled during a provisional labeling stage after the first NER. The completed synonym dictionary and translation process can be applied to new data, improving prediction accuracy.

The original dataset comprised 30,119 raw sentences from BoQ and BIM, from which 3,610 sentences were selected for training and testing. The dataset was split into 80% training data and 20% test data to develop the NER model. This structured methodology ensures high-quality entity extraction while enhancing the efficiency of text processing related to construction.

3.5. Training BERT-based NER model

The training process of the BERT-based NER model used in this study consists of the following steps:

1. BERT Embedding Transformation of Input Sentences

The input sentence undergoes tokenization and is transformed through the embedding layer of the BERT

Table 3. Imbalanced data treatment for NER

Division	Language	Number of Sentence (Before)	Number of Sentence (After)
PJT1 (BoQ → CBS)	Korean, English	1,160	1,160
PJT2 (BoQ)	Korean, English	122	366
PJT3 (BoQ)	English	51	153
PJT4 (BoQ)	English	92	276
PJT2 (BIM)	Korean, English	8,474	644
PJT3 (BIM)	English	20,220	1,011
Total		30,119	3,610

model. A tokenizer is used to split sentences into subword units, followed by the addition of a classification token (CLS) and a sentence separator token (SEP). The transformed sentence consists of Token Embeddings (vectors representing each word split into WordPiece tokens), Segment Embeddings (vectors distinguishing different segments within a sentence), and Position Embeddings (vectors incorporating positional information to reflect contextual meaning), which are processed into a 768-dimensional input vector and fed into BERT. At this stage, each token is labeled using the BIO tagging scheme. 'B-X' means the beginning of an entity, 'I-X' means intermediate or last token in an entity, and 'O' means outside any entity class.

2. Passing the Embeddings through the BERT Transformer Layers

BERT consists of a deep neural network with 12 Transformer blocks, each containing multi-head self-attention and feed-forward layers that hierarchically apply non-linear transformations to the input data.

The attention operation in each Transformer block is defined as follows:

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, and V represent the Query, Key, and Value matrices, and d_k is the scaling factor, typically equal to the matrix dimension.

$$FFN\left(x\right)=\ W_2\cdot f\left(W_1\cdot x+\ b_1\right)+\ b_2. \tag{2}$$

The feed-forward neural network (FFNN) in each layer consists of two linear transformations (W_1, W_2), biases (b_1, b_2), and an activation function $f(x)$. After passing through all Transformer layers, the final hidden state (H) of BERT is generated.

3. Adding a Dense Classification Layer for Fine-Tuning

In the NER task, each word must be assigned a specific entity tag. To achieve this, the H from BERT is used for classification by adding a dense output layer that predicts entity tags:

$$P=softmax\left(WH+b\right), \tag{3}$$

where W represents the trainable weight matrix, b is the bias vector, and the softmax function produces probability scores for each token's entity class.

4. Optimization

The Adam optimizer is employed for model optimization. Adam extends stochastic gradient descent (SGD) by incorporating momentum-based learning rate adjustments to improve convergence. The weight update rule for Adam is given by:

$$m_t=\beta_1 m_{\{t-1\}}+\left(1-\beta_1\right)g_t; \tag{4}$$

$$v_t=\beta_2 v_{t-1}+\left(1-\beta_2\right)g_t^2; \tag{5}$$

$$\theta t=\theta_{t-1}-\frac{\alpha}{\sqrt{\{v_t\}}+\epsilon}m_t, \tag{6}$$

where g_t, is the current gradient, m_t represents the momentum term, and v_t is the adaptive learning rate adjustment term.

5. Loss Function Definition

For training the NER model, Sparse Categorical Crossentropy is used as the loss function, as it is well-suited for multi-class classification tasks where labels are represented as integers instead of one-hot vectors:

$$L=-\sum_i y_i\log\left(\widehat{y_i}\right), \tag{7}$$

where y_i, is the true label and $\widehat{y_i}$ is the predicted probability.

6. Model Compilation and Training Execution

Once the loss function is defined, the BERT-based NER model is trained using a mini-batch gradient descent approach with a batch size of 32 over 10 epochs.

Weight updates are computed using the Adam optimizer, following the gradient of the loss function:

$$\theta_{t+1}=\theta_t-\eta\nabla L\left(\theta\right), \tag{8}$$

here, η is the learning rate and $\nabla L\left(\theta\right)$ represents the gradient of the loss function.

The final hidden state (H) of BERT undergoes a linear transformation to generate predictions. During evaluation and inference, the highest probability class is selected using an argmax operation. The predicted entity tags are directly utilized in the loss function without additional normalization.

3.6. Evaluation of BERT-based NER model

To assess the performance of the NER model, precision, recall, and F1-score were calculated using an F1-score callback to monitor model performance. The F1-score, which represents the harmonic mean of precision and recall, is computed as follows:

$$F1=2\times\frac{Precision\times Recall}{Precision+Recall}, \tag{9}$$

here:

- $Precision=\frac{TP}{TP+\ FP}$;
- $Recall=\frac{TP}{TP+FN}$;
- $TP\left(True\ Positives\right)$: The number of correctly predicted positive instances;
- $FP\left(False\ Positives\right)$: The number of incorrectly predicted positive instances;
- $FN\left(False\ Negatives\right)$: The number of actual positive instances that were misclassified as negative.

The F1-score results for the NER model in this study, computed based on the above formula, are presented in Table 4.

The "NAM" category achieved a high F1-score, likely due to its consistent position at the beginning of sentences and the presence of the "/" delimiter. Similarly, the "MAT" category achieved a high score, as it comprises

Table 4. F1 score of BERT-NER

	BERT-NER			Bi-LSTM-CRF			GPT		
Label	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
NAM	0.96	0.97	0.96	0.91	0.99	0.95	0.98	0.86	0.92
MAT	0.97	0.96	0.97	0.80	0.88	0.84	0.98	0.73	0.84
SIZ	0.98	0.96	0.97	0.98	0.81	0.89	0.96	0.96	0.96
THK	0.97	0.95	0.96	0.79	0.87	0.83	0.87	0.89	0.88
DIA	0.98	0.96	0.97	0.93	0.97	0.95	0.95	0.95	0.95
LEN	0.98	0.94	0.96	0.75	0.87	0.80	0.92	0.9	0.91
TYP	0.94	0.97	0.95	0.78	0.90	0.83	0.91	0.96	0.93
ETC	0.92	0.73	0.81	0.76	0.88	0.81	0.84	0.84	0.84

a relatively small number of material types (e.g., Concrete, Steel, SM355, SS275), with a substantial amount of training data. Additionally, "SIZ", "THK", and "DIA" categories performed well due to their frequent association with specific textual patterns, such as "x", "THK", "D", and "Ø". However, the performance of the "ETC" category was relatively low, likely due to low recall (0.73) despite reasonable precision (0.92), indicating difficulty in capturing miscellaneous text variations with the unstructured, non-repetitive, and inconsistent nature of the attribute information, which labels specific details (scope of application, number of repetitions, application conditions, etc.).

The Bi-LSTM-CRF approach showed moderate performance with F1 scores between 0.80–0.95. While achieving competitive results for NAM (F1 = 0.95) and DIA (F1 = 0.95), the model demonstrated significant performance drops for technical specifications including LEN (F1 = 0.80) and TYP (F1 = 0.83). The model exhibited consistent precision-recall imbalances, with generally lower precision values across most categories compared to BERT-NER.

GPT showed the most variable performance with F1 scores ranging from 0.84 to 0.96. Interestingly, GPT achieved the highest precision scores for several categories: NAM (0.98), MAT (0.98), and competitive performance for SIZ (0.96). However, this high precision came at the cost of lower recall, particularly evident in MAT (recall = 0.73) and NAM (recall = 0.86). The pattern suggests GPT adopts a conservative prediction strategy, making fewer but more confident predictions.

The precision-recall trade-off is most pronounced in GPT (gpt-4o-mini), which achieves high precision through conservative prediction but sacrifices recall. BERT-NER maintains the best precision-recall balance across most categories. The consistent superior performance of BERT-NER (average F1 = 0.94) over Bi-LSTM-CRF (average F1 = 0.87) and GPT (average F1 = 0.89) demonstrates the effectiveness of contextual embeddings for construction domain NER tasks.

The extracted entity data from BIM parameters and BoQ descriptions based on the final NER model are illustrated in the following figures.

Referring to Figure 5, it is evident that BIM library names and BoQ item names exhibit different structural patterns. BIM library names primarily consist of Element, Type, and Material, making BoQ item names a more suitable reference for annotating 'NAM' entities. This ratio-

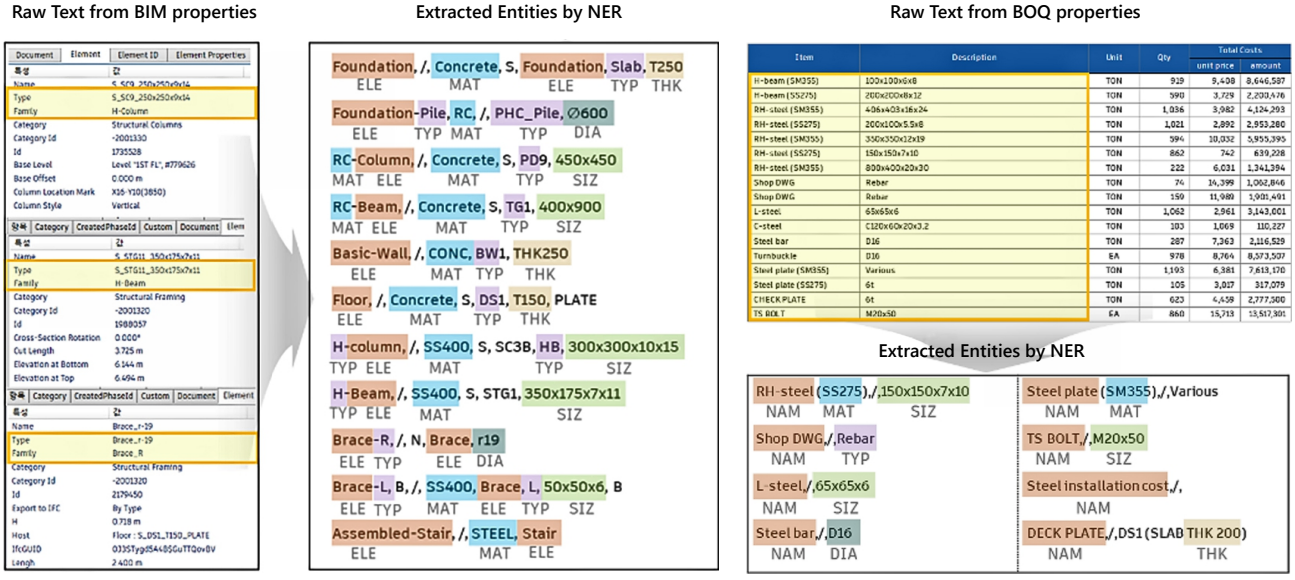

Figure 5. NER results from construction documents: (left) Entity extraction result from BIM parameters; (right) Entity extraction result from BoQ description

nale aligns with the approach taken in the 'Metadata for Ontology' section, where CBS was constructed based on BoQ. Furthermore, since the entities extracted from BoQ and BIM have differing naming conventions, a similarity search is conducted in the following section to link them based on CBS names.

3.7. Semantic text similarity

While the previous section outlined entity extraction from BoQ and BIM sources, these datasets remain independent corpora rather than unified datasets. Furthermore, since the BIM library name is structured based on element types, it is challenging to achieve a one-to-one direct mapping with the BoQ, which does not contain element information. This issue is also a common difficulty encountered by practitioners when comparing quantities between BIM and BoQ.

To address this challenge, this section integrates BoQ entities and BIM entities into a unified dataset using STS analysis, with CBS as the central reference point (Figure 6). The applied STS process consists of the following steps:

1. Data Pre-processing

Initially, text labeled as 'O' (indicating non-meaningful tokens) was removed before performing similarity matching. Preliminary tests revealed that the structural differences between BoQ and BIM data components resulted in an initial accuracy of only 0.72. Since BERT embeddings are influenced by word order, the placement of BIM element entities and type entities before material entities may have affected this issue. To address this, entity labels were automatically reordered in the sequence: NAM → MAT → TYP → SIZ → ELE → ETC, and the following similarity matching algorithm was applied.

Additionally, the CBS items were classified into two categories: those with and without BIM modeling. If this step were omitted, BIM text such as "RC-Beam, Concrete, 400×600" would have been matched to the CBS category "Concrete Casting". However, after the reclassification, it was correctly mapped to "Ready-Mix Concrete".

2. Similarity search

$$e_{BoQ},\ e_{BIM},\ e_{CBS} \in R^d, \tag{10}$$

here:

- e_{BoQ}, e_{BIM}, and e_{CBS} are text embedding vectors of each BoQ entities as query, BIM entities as query, and CBS entities as database.
- R^d is a set of real numbers, and d is the number of embedding dimensions.

$$sim\left(e_{BoQ}, e_{CBS}\right) = \frac{\sum_i \left(e_{BoQ,i} \cdot e_{CBS,i}\right)}{\| \sum_i e_{BoQ,i} \| \| \sum_i e_{CBS,i} \|}; \tag{11}$$

$$sim\left(e_{BIM}, e_{CBS}\right) = \frac{\sum_i \left(e_{BIM,i} \cdot e_{CBS,i}\right)}{\| \sum_i e_{BIM,i} \| \| \sum_i e_{CBS,i} \|}. \tag{12}$$

The texts composed of entities from BoQ and BIM respectively, are calculated for similarity with entities from CBS. The STS results for the ontology development followed a systematic approach utilizing expert-validated ground truth data as the foundation for automated entity extraction and evaluation. Three BIM specialists with 6, 10, and 15 years of professional experience pre-established CBS-BoQ mapping datasets and EBS-BIM parameter matching data, which served as reference materials for subsequent NER and STS evaluation. Following text similarity matching, the final accuracy reached approximately 94%.

The accuracy results for the top 11 CBS items with high matching rates among 2,450 total query items from BIM and BoQ are presented in Table 5. Cases such as "BH-steel, Steel-plate, Concrete-casting, Micro-pile-manufacturing-and-installation, Steel-stair-installation" achieved 100% accuracy in CBS item matching, while "C-steel, L-steel, Pile-drilling-and-driving" demonstrated accuracies of 71%, 53%, and 0%, respectively. "C-steel, Pile-drilling-and-driving" items were matched to different CBS items due to ELE class entities being positioned at the beginning of sentences where NAM class entities are typically located,

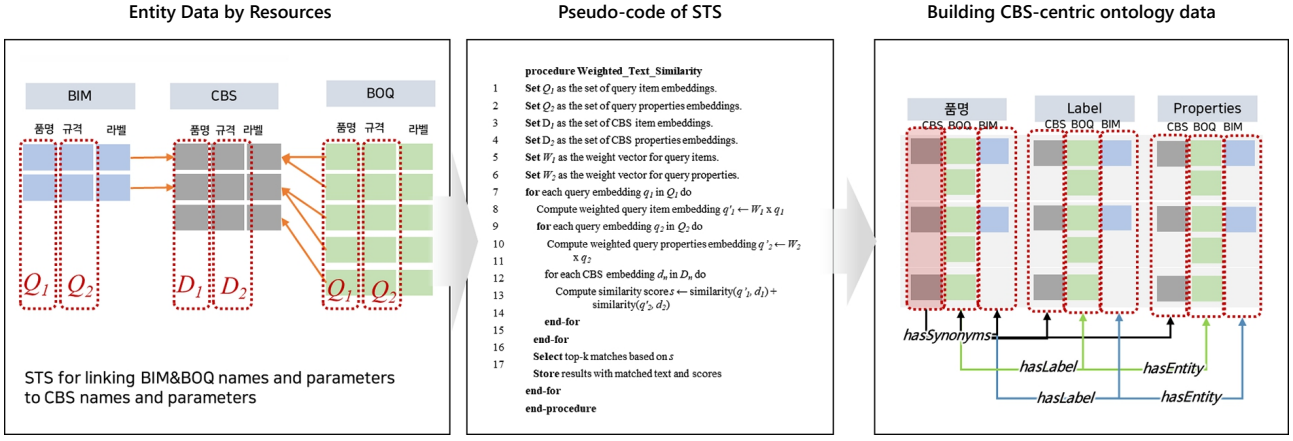

Figure 6. STS process for CBS-centered entity data integration showing: (1) entity extraction from BIM, BoQ, and CBS sources; (2) similarity matching algorithm implementation; (3) resulting unified ontology structure with standardized entity relationships and properties

Table 5. Accuracy of STS Results by CBS items

CBS item	Right Query	Wrong Query	Accuracy
PHC-pile	122	1	99%
BH-steel	639	0	100%
C-steel	10	4	71%
Steel-plate	10	0	100%
Concrete-casting	236	0	100%
RH-steel	484	14	97%
L-steel	40	36	53%
Micro-pile-manufacturing-and-installation	10	0	100%
Pile-drilling-and-driving	0	37	0%
Steel-stair-installation	5	0	100%

which can be addressed by modifying ELE entity positioning. For "L-steel" cases, most queries involved incorrect matching of "H-Beam", which was resolved by converting "H-Beam" to "H-steel" through synonym dictionary implementation.

In the third and seventh items of Table 6, the BIM element entity was repositioned to the end of the sentence according to the 'Data Pre-processing' step. This adjustment was necessary to prevent a decline in CBS text matching accuracy when element names appeared at the beginning of the text. The seventh item exhibited high similarity between "RC" and "RH-Steel", indicating a need to refine the data pre-processing step by automatically converting "RC" to "Concrete" in specific contexts.

Through the STS process, previously independent BoQ and BIM entities were successfully integrated with CBS as the central reference. The NAM entity extracted from BoQ and BIM was consolidated as a CBS synonym, while other extracted entities were incorporated as CBS entities.

4. Result: Integration of project documents for an extended ontology dataset

In the previous section, the extracted entities were integrated as ontology components. The model development database was constructed using Excel-based management, with Entity-Class mapping sheets systematically created through NER to structurally organize entity relationships and semantic connections across construction datasets.

The core ontology structure encompasses eight primary entity classes (NAM, ELE, MAT, SIZ, THK, DIA, TYP, ETC) while comprehensively covering commonly used construction parameter categories. Manual labeling of BoQ from representative projects provided the foundation for CBS development, with NAM class entities serving as the central reference point for ontological relationships.

The developed ontology successfully standardized 1,160 BoQ items from PJT1 into CBS through expert-validated mapping procedures. A total of 2,450 items from PJT2-4 BoQ and BIM datasets achieved F1 scores ranging from 0.81 to 0.97 after entity and class extraction followed by cross-referencing through mapping sheets. The extracted entities by class were mapped to CBS with 94% accuracy through STS and systematically documented to establish semantic connections between heterogeneous construction data sources. This framework includes 23 synonyms for Korean-English construction term mapping and 18 synonyms for English-English construction abbreviation mapping, enabling terminology standardization across diverse linguistic contexts. Cross-system validation for BIM and BoQ entity matching achieved 94% accuracy when compared against expert-prepared ground truth datasets. This accuracy increased by approximately 4% following synonym dictionary construction. This validation process confirmed the semantic consistency and practical applicability of the ontological framework for real-world construction data integration scenarios.

Two key findings emerged during the STS-based data mapping process: 1) When calculating text similarity with Name Entity and property Entity weighted at ratios of 3:7, 5:5, 6:4, and 8:2, the 8:2 ratio yielded the lowest accuracy at approximately 45%, while the 6:4 ratio achieved the highest accuracy at 94%. 2) Although threshold values were examined by reviewing correct answers based on text similarity values, correct data distributed within a similarity range of 0.08–0.74, while incorrect data distributed within 0.39–0.71, confirming that threshold-based classification for answer accuracy was not feasible.

Figure 7 illustrates a conceptual example of an ontology for foundation work, steel framework, and reinforced concrete (RC) work, defining the hierarchical structure of various entities and CBS-specific parameters:

- Pile: Represents an entity that defines synonyms, pile material, pile diameter, and structural components.
- H-Steel: Represents an entity that defines synonyms, H-section steel material, H-section steel size, and structural components.

Table 6. Example of the STS results

Index	Query (BoQ, BIM)	CBS	Results
1	PHC-Pile / RC, Ø500, Foundation_Pile	PHC-pile / D500	OK
2	Concrete, MF1, THK.700, Foundation_Slab	Reinforced-concrete-casting / Foundation	OK
3	CONC, THK200, Basic_Wall	Concrete-casting / Wall	OK
4	H-Steel / SM 355, H, 582x300x12x17, Structural_Beam	RH-steel / SM355, 582x300x12x17	OK
5	Brace-L, SS400, L, 50x50x6, B, Brace	L-steel / SS275, 50x50x 4	OK
6	Assembled-Stair / Steel, Stair	Steel-stair-installation / 1300x2620x1160H	OK
7	RC, 400x1650, Structural_Beam	RH-steel / SHN275, 400x400x13x21	NG

- Painting: Represents an entity that defines synonyms, paint type, and additional descriptions. Since painting is not modeled in BIM, it does not contain an ELE (Element) entity.
- Ready-Mix Concrete: Represents an entity that defines synonyms, concrete material, concrete thickness, and structural components.

This example serves as a prototype for how the proposed ontology structures CBS and EBS data. For instance, "hasMAT" in BoQ and "hasMAT" in BIM can be linked to facilitate comparisons of quantities and construction costs.

This study reduced manual effort and improved efficiency while maintaining high accuracy. A comparative analysis of the manual annotation process and the proposed automated pipeline highlights substantial improvements (Table 7).

For a total of 3,610 sentences requiring annotation, the manual annotation process takes approximately 28 seconds per sentence (appx.28hrs in total), while post-correction of automatically generated labels takes about 9 seconds per sentence (appx. 3 hrs in total). Therefore, the semi-automated approach demonstrates a substantial improvement in efficiency, reducing the amount of manually labeled data by approximately 83% and the processing time by approximately 70%, resulting in a time reduction of 90% and an accuracy improvement of 9% compared to fully manual labeling methods for a new NER model development.

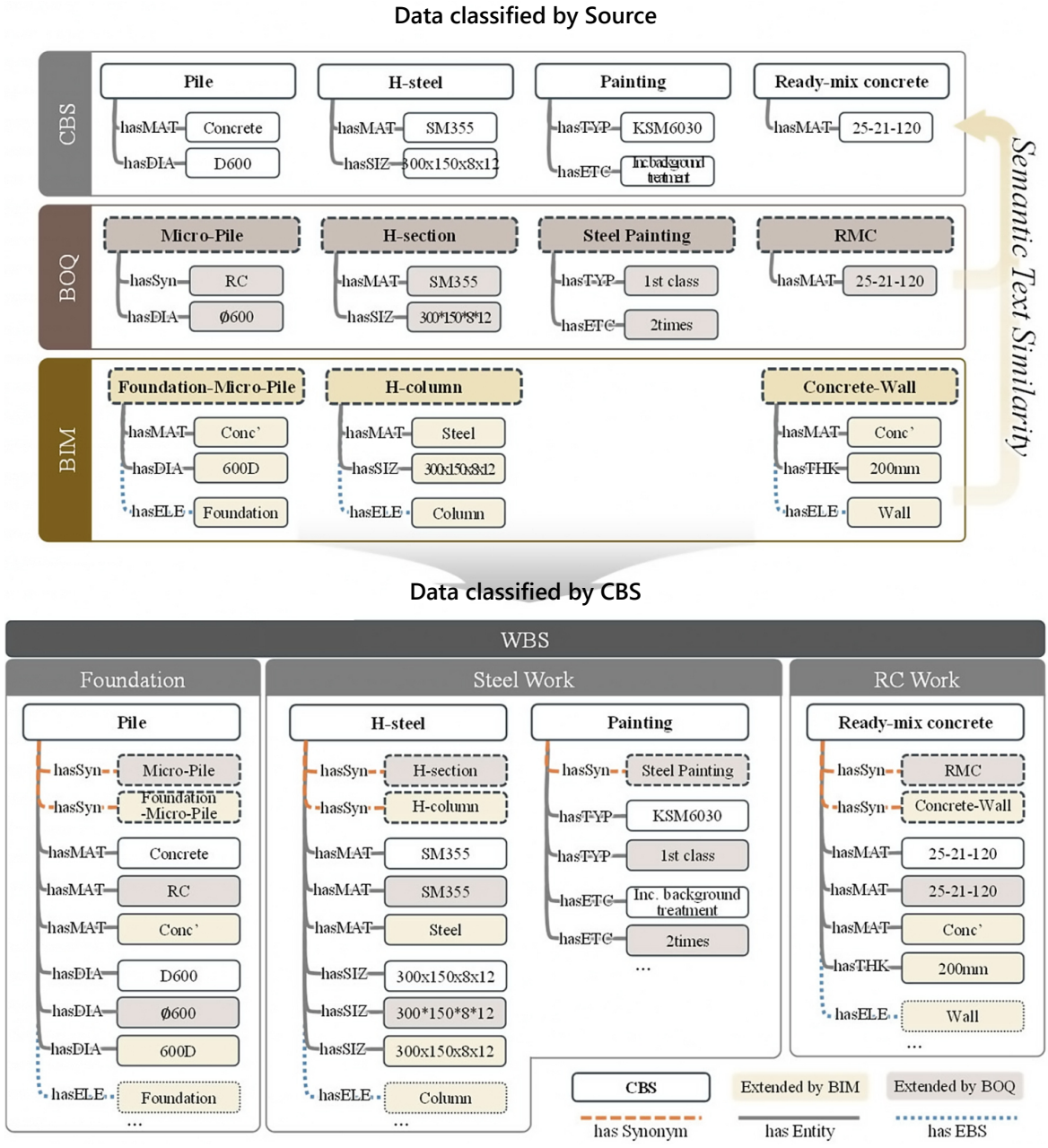

Figure 7. Example of Ontology Result Data transforming heterogeneous source data (CBS, BoQ, BIM) into unified CBS structure with standardized entity-property relationships and cross-system synonym mapping capabilities

Table 7. Results of STS

	Human labor	BERT-based NER	Comparison
Speed	28 seconds/case	9 seconds/case	About 3 times
Time	About 28 hours	About 3 hours	About 90% or less
Performance	Average 85% (human error standard)	94% on average (based on F1 score)	About 9% increase

Beyond efficiency gains, the integration of ontology-based structured metadata further optimized data classification and interoperability between BoQ description and BIM parameters. The alignment of extracted entities through semantic text similarity (STS) reduced discrepancies in naming conventions, facilitating consistent data retrieval and analysis. The results confirm that BoQ item names serve as a more suitable reference than BIM library names for defining standardized entities.

This study also highlights critical considerations for applying NER to construction domain text. Unlike general-purpose entity extraction, BoQ and BIM data exhibit domain-specific abbreviations, non-standardized terminology, and contextual ambiguity, requiring tailored preprocessing and annotation strategies. Our findings suggest that the proposed approach can be adapted to other domains with appropriate modifications.

5. Conclusions

This study proposed an automated NER and ontology-based integration framework for extracting and linking entities from BoQ descriptions and BIM parameters with minimal human intervention.

The current technical framework establishes a foundation using software architectures including TensorFlow for deep learning operations and Hugging Face Transformers for BERT model deployment. Separate translation processing through Google Translate API demonstrated improved accuracy compared to the BERT-multilingual models when validated with the Korean-English dataset.

However, deployment challenges arise from the substantial computational requirements and complex dependency management inherent in BERT-based systems, necessitating model compression and minimization through Docker containerization and cloud deployment via Kubernetes orchestration for enterprise-scale applications.

This study processed over 30,000 sentences using CSV file-based data management through the Pandas library. To address scalability constraints in future enterprise deployment, more robust solutions are required, such as Parquet for enhanced processing speed or relational database systems (e.g., SQLite, PostgreSQL) for concurrent user environments and data integrity assurance.

The developed ontology framework can enrich the contextual understanding of construction data and address the industry's growing demand for integrated project management platforms. When synchronized in real-time across heterogeneous software environments via API endpoints and integrated with work breakdown structures (WBS), cost information, and quantity data, the ontology framework can support seamless project management integration across design, procurement, and construction phases.

Design applications benefit from parametric rule validation and standardized vocabularies, while procurement processes gain automated quantity verification and supplier integration capabilities. Construction phase operations achieve enhanced progress tracking and quality control through consistent data categorization and reporting frameworks. Additionally, project progress monitoring becomes feasible through data integration between design BIM, construction BoQ, and as-built BIM. This advancement ultimately promises improved efficiency, accuracy, and project management effectiveness through comprehensive information integration throughout the construction lifecycle.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00277939).

This research was conducted with the support of the "National R&D Project for Smart Construction Technology (No. RS-2020-KA156875)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation.

Funding

This work was supported by the National Research Foundation of Korea (NRF) under Grant RS-2023-00277939; and Korea Agency for Infrastructure Technology Advancement (KAIA) under Grant RS-2020-KA156875.

Disclosure statement

All Authors confirm there is no competing financial, professional, or personal interests from other parties.

References

Arzideh, K., Schäfer, H., Allende-Cid, H., Baldini, G., Hilser, T., Idrissi-Yaghir, A., Laue, K., Chakraborty, N., Doll, N., Antweiler, D., Klug, K., Beck, N., Giesselbach, S., Friedrich, C. M., Nensa, F., Schuler, M., & Hosch, R. (2025). From BERT to generative AI – Comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in

unstructured medical reports. *Computers in Biology and Medicine*, *195*, Article 110665. https://doi.org/10.1016/j.compbiomed.2025.110665

Beetz, J., van Leeuwen, J., & de Vries, B. (2009). IfcOWL: A case of transforming EXPRESS schemas into ontologies. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *23*(1), 89–101. https://doi.org/10.1017/S0890060409000122

Cho, H., & Lee, H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, *20*, Article 735. https://doi.org/10.1186/s12859-019-3321-4

Gruber, T. (1993). Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino, & R. Poli (Eds.), *Formal ontology in conceptual analysis and knowledge representation*. Kluwer Academic Publishers.

Halmetoja, E. (2022). The role of digital twins and their application for the built environment. In M. Bolpagni, R. Gavina, & D. Ribeiro (Eds.), *Industry 4.0 for the built environment*: *Vol. 20*. *Structural integrity* (pp. 415–442). Springer, Cham. https://doi.org/10.1007/978-3-030-82430-3_18

Jagannathan, M., Roy, D., & Delhi, V. S. K. (2022). Application of NLP-based topic modeling to analyse unstructured text data in annual reports of construction contracting companies. *CSI Transactions on ICT*, *10*(2), 97–106. https://doi.org/10.1007/s40012-022-00355-w

Jeon, K., Lee, G., Yang, S., & Jeong, H. D. (2022). Named entity recognition of building construction defect information from text with linguistic noise. *Automation in Construction*, *143*, Article 104543. https://doi.org/10.1016/j.autcon.2022.104543

Jeong, D. W., Park, J. H., Seo, J. J., Shin, W. H., & Jang, H. Y. (2024). The progress and prospective advancements of the national digital twin pilot project. *Journal of Korean Society for Geospatial Information Science*, *32*(3), 51–63. https://doi.org/10.7319/kogsis.2024.32.3.051

Keshavarz, H., Vagena, Z., Kouki, P., Fountalis, I., Mabrouki, M., Belaweid, A., & Vasiloglou, N. (2022). Named entity recognition in long documents: An end-to-end case study in the legal domain. In *2022 IEEE International Conference on Big Data* (*Big Data*), Osaka, Japan. IEEE. https://doi.org/10.1109/BigData55660.2022.10020873

Kim, S.-Y., Lee, H.-H., Choi, E.-S., & Go, J.-U. (2020). A case study on the construction of 3D geo-spatial information for digital twin implementation. *Journal of the Korean Association of Geographic Information Studies*, *23*(3), 146–160.

Kuiper, I., & Duffield, C. (2018). *Describing structural configurations towards identifying and establishing theoretical foundations for the exploration and understanding of building information modelling* (*BIM*). Department of Infrastructure Engineering, The University of Melbourne.

Lê, N. C., Nguyen, N.-Y., & Trinh, A. D. (2019). On the Vietnamese name entity recognition: A Deep Learning Method approach. In *2020 RIVF International Conference on Computing and Communication Technologies* (*RIVF*), Ho Chi Minh City, Vietnam. IEEE. https://doi.org/10.1109/RIVF48685.2020.9140754

Lester, B., Pressel, D., Hemmeter, A., Choudhury, S. R., & Bangalore, S. (2020). *Multiple word embeddings for increased diversity of representation*. arXiv. https://doi.org/10.48550/arXiv.2009.14394

Li, S., Wang, J., & Xu, Z. (2024). Automated compliance checking for BIM models based on Chinese-NLP and knowledge graph: An integrative conceptual framework. *Engineering, Construction and Architectural Management*, *32*(6), 3832–3856. https://doi.org/10.1108/ECAM-10-2023-1037

Luo, Y., Xiao, F., & Hai, Z. (2019). *Hierarchical contextualized representation for named entity recognition*. ArXiv. https://doi.org/10.48550/arXiv.1911.02257

Luoma, J., & Pyysalo, S. (2020). Exploring cross-sentence contexts for Named Entity Recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 904–914), Barcelona, Spain. https://doi.org/10.18653/v1/2020.coling-main.78

Na, Y. G., & Kim, J. Y. (2024). Metadata design for interoperability of digital land information. *Journal of the Korean Cadastre Information Association*, *26*(3), 133–145. https://doi.org/10.46416/JKCIA.2024.12.26.3.133

Pakhale, K. (2023). *Comprehensive overview of named entity recognition*: *Models*, *domain-specific applications and challenges*. arXiv. https://doi.org/10.48550/arXiv.2309.14084

Rasmussen, M. H., Lefrançois, M., Schneider, G. F., & Pauwels, P. (2021). BOT: The building topology ontology of the W3C linked building data group. *Semantic Web*, *12*(1), 143–161. https://doi.org/10.3233/SW-200385

Sammet, J., & Krestel, R. (2023, September). Domain-specific keyword extraction using BERT. In S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch, & A. Salgado (Eds.), *Proceedings of the 4th Conference on Language*, *Data and Knowledge* (pp. 659–665), Vienna, Austria.

Taher, E., Hoseini, S. A., & Shamsfard, M. (2020). *Beheshti-NER*: *Persian named entity recognition using BERT*. arXiv. https://doi.org/10.48550/arXiv.2003.08875

Taillé, B., Guigue, V., Gallinari, P., & Paribas, B. (2019). Une Étude Empirique de la Capacité de Généralisation des Plongements de Mots Contextuels en Extraction d'Entités. In *Conférence Nationale d'Intelligence Artificielle Année 2019*.

Tang, S., Zhang, C., Hao, J., & Guo, F. (2022). A framework for BIM, BAS, and IoT data exchange using semantic web technologies. In *Construction Research Congress 2022*. ASCE. https://doi.org/10.1061/9780784483961.098

Wu, S., Shen, Q., Deng, Y., & Cheng, J. (2019). Natural-language-based intelligent retrieval engine for BIM object database. *Computers in Industry*, *108*, 73–88. https://doi.org/10.1016/j.compind.2019.02.016

Wu, C., Wang, X., Wu, P., Wang, J., Jiang, R., Chen, M., & Swapan, M. (2021). Hybrid deep learning model for automating constraint modelling in advanced working packaging. *Automation in Construction*, *127*, Article 103733. https://doi.org/10.1016/j.autcon.2021.103733

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022a). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, *134*, Article 104059. https://doi.org/10.1016/j.autcon.2021.104059

Wu, L.-T., Lin, J.-R., Leng, S., Li, J.-L., & Hu, Z.-Z. (2022b). Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. *Automation in Construction*, *135*, Article 104108. https://doi.org/10.1016/j.autcon.2021.104108

Wu, D., Yang, J., & Wang, K. (2024). Exploring the reversal curse and other deductive logical reasoning in BERT and GPT-based large language models. *Patterns*, *5*(9), Article 101030. https://doi.org/10.1016/j.patter.2024.101030

Xie, S. (2024). Research on Named Entity Recognition Method based on BERT model. In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications* (*BigDataService*) (pp. 92–96), Shanghai, China. IEEE. https://doi.org/10.1109/BigDataService62917.2024.00020

Xu, X., & Cai, H. (2021). Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure. *Advanced Engineering Informatics*, *48*, Article 101288. https://doi.org/10.1016/j.aei.2021.101288

Yang, G., & Xu, H. (2020). A residual BiLSTM model for named entity recognition. *IEEE Access*, *8*, 227710–227718. https://doi.org/10.1109/ACCESS.2020.3046253

Yin, M., Tang, L., Webster, C., Li, J., Li, H., Wu, Z., & Cheng, R. C. (2023). Two-stage Text-to-BIMQL semantic parsing for building information model extraction using graph neural networks. *Automation in Construction*, *152*, Article 104902. https://doi.org/10.1016/j.autcon.2023.104902

Yin, M., Tang, L., Webster, C., Yi, X., Ying, H., & Wen, Y. (2024). A deep natural language processing-based method for ontology learning of project-specific properties from building information models. *Computer-Aided Civil and Infrastructure Engineering*, *39*(1), 20–45. https://doi.org/10.1111/mice.13013

Yun, J., & Kim, J. (2022). *An analysis of research and standardization trends on digital twin*. Society for Standards Certification and Safety. https://doi.org/10.34139/JSCS.2022.12.1.31

Zhang, Y., & Zhang, H. (2023). FinBERT–MRC: financial named entity recognition using BERT under the machine reading comprehension paradigm. *Neural Processing Letters*, *55*(6), 7393–7413. https://doi.org/10.1007/s11063-023-11266-5

Zhang, Q., Xue, C., Su, X., Zhou, P., Wang, X., & Zhang, J. (2023). Named entity recognition for Chinese construction documents based on conditional random field. *Frontiers of Engineering Management*, *10*(2), 237–249. https://doi.org/10.1007/s42524-021-0179-8