

Transformer-based Approach for Automated Context-aware IFC-regulation Semantic Information Alignment

Ruichuan Zhang[a]; and Nora El-Gohary[b]

[a] Graduate Student, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, United States. E-mail: rzhang65@illinois.edu.

[b] Associate Professor, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, United States (corresponding author). E-mail: gohary@illinois.edu; Tel: +1-217-333-6620.

Abstract

One of the main challenges of automated compliance checking systems is aligning the semantics of the building information models (BIM), in Industry Foundation Classes (IFC) format, and the semantics of the regulations, in natural language, to allow for checking the compliance of the BIM with the regulations. Existing information alignment methods typically require intensive manual effort and their ability to deal with the complex regulatory concepts in the regulations is limited. To address this gap, this paper proposes a deep learning method for IFC-regulation semantic information alignment. The proposed method uses a relation classification model to relate and align the IFC and regulatory concepts. The method uses a transformer-based model and leverages the definitions of the concepts and an IFC knowledge graph to provide additional contextual information and knowledge for improved classification and alignment. The proposed method was evaluated on IFC concepts from IFC 4 and regulatory concepts from different building codes and standards. The experimental results showed good information alignment performance.

Keywords: Information alignment; Automated code checking; Building codes; Building information modeling; Industry Foundation Classes; Deep learning; Transformers.

1 Introduction

Building designs are governed by a wide range of regulations and requirements in the architecture, engineering, and construction (AEC) domain, such as building codes, standards, and specifications. To improve regulatory and contract compliance, as well as project efficiency, various automated compliance checking (ACC) systems have been developed with the aim of automating – fully or partially – the process of checking the compliance of building designs, captured in building information models (BIM), with applicable regulations and requirements. However, a bottleneck in the ACC process is bridging the semantic gap between the BIM [commonly represented using the Industry Foundation Classes (IFC) schema] and the regulations (expressed in natural language such as English) [1-3]. Before conducting the compliance checking, it is essential to align the semantic representations and terminology of the IFC to that of the natural-language regulations.

In most of the existing ACC systems, such information alignment is conducted in a highly manual way, through hardcoding (e.g., using modeling or query languages), ontology- or dictionary-based matching, or searching methods. For example, the buildingSMART Data Dictionary (bSDD) [4], an online service that provides access to classifications (e.g., Uniclass) related to the AEC domain, can be used to facilitate the matching of regulatory concepts to their corresponding IFC concepts (e.g., IFC entities, properties, or enumerated property values). These methods require intensive manual effort and are by nature rigid and difficult to generalize [3, 5-6]. Also, they are less capable to deal with semantically or syntactically complex regulatory concepts. For example, many single-word regulatory concepts can be directly matched to IFC concepts (e.g., match "beam" to "IfcBeam" or "IfcBeamTypeEnum – Beam"); however, it is difficult to match multi-word, phrasal, or clausal regulatory concepts directly to any of the IFC concepts [e.g., "membrane-covered frame structure" and "intended to be occupied as a residence" in the International Building Code (IBC) [7]]. There is, thus, a need for an automated, and meanwhile flexible and generalizable, method for IFC-regulation semantic information alignment for supporting fully automated ACC.

Towards addressing this need, the most recent efforts that focused on IFC-regulation semantic information alignment have explored the use of machine learning to facilitate such automation. Instead of relying on hardcoding or handcrafted rules, these efforts use machine learning models to automatically learn the underlying semantic and syntactic patterns of the regulatory text and IFC data to help in the alignment. Many of these efforts focused on augmenting the BIM models with additional attributes and relationships to support the alignment for ACC (e.g., [9-11]), while other efforts focused on directly aligning the regulatory and IFC concepts (e.g., [8]). For example, Wang et al. [11] modeled IFC-based building designs as graphs and used graph neural networks (GNN) to classify the rooms in the IFC models into nine predefined types based on manually constructed node and edge features and augment the models with the classified types. Zhou and El-Gohary [8] leveraged word and concept semantic representations learned using the word2vec algorithm and the graph structures of the IFC-based building designs to align concepts from the International Energy Conservation Code (IECC) and energy specifications to their corresponding IFC concepts. However, despite their importance, both groups of efforts still lack in flexibility and adaptability and might not allow successful implementation across different BIMs and different types of regulatory documents (e.g., building code versus energy code) due to two reasons. First, they rely on contextless features (e.g., the word2vec representations), which have limited ability to capture the semantic and syntactic dependencies of IFC and text data. Second, they have not exploited the contextual information and knowledge in both the IFC schema and the regulatory documents, which can potentially provide additional semantic information for aligning IFC and regulatory concepts.

To address this need, this paper proposes a transformer-based method to align regulatory concepts in the requirements with the IFC concepts in the IFC schema for supporting downstream ACC information matching and compliance reasoning processes. The proposed method uses a relation classification model to classify each pair of IFC-regulatory concepts as semantically related or not. The method utilizes the natural-language definitions of the concepts and an IFC knowledge graph to provide additional contextual information and knowledge for the classification. It also leverages semantic and syntactic patterns learned in pretrained transformer-based language models, as well as domain-specific semantic and syntactic patterns learned using transfer learning strategies. The proposed method was tested on IFC concepts and definitions from IFC Version 4, and regulatory concepts and definitions from three different types of regulatory documents including IBC, IECC, and Americans with Disabilities Act Standards for Accessible Design (ADA Standards), and an average precision of 84.3%, recall of 83.3%, and F1 measure of 83.8% in alignment was achieved.

2 Background

2.1 Deep learning in text and knowledge analytics

Deep learning methods use deep neural networks to capture multiple levels of information representations from large-scale data [12]. Deep learning methods have been used in solving various text analytics tasks, such as information extraction [e.g., bidirectional long short-term memory (LSTM) and conditional random fields for extracting named entities [13]], semantic and syntactic analysis (e.g., bidirectional LSTM for dependency parsing and part-of-speech tagging [14]), and machine translation [e.g., sequence-to-sequence recurrent neural network (RNN) model for machine translation [15]]. Deep learning methods have also been used in solving various knowledge analytics tasks (especially the ones related to knowledge graphs), such as relation analysis (e.g., relation adversarial network [16], relation attention network [17]), knowledge graph embedding learning (e.g., GNN and negative sampling [18], GNN with contrastive learning [19]), and knowledge graph-based question answering and recommendation (e.g., LSTM- and attention-based method [20] and GNN- and attention-based method [21]).

A number of research efforts have focused on deep learning-based methods to solve text or knowledge analytics problems in the AEC domain. For example, Pan and Zhang [22] developed RNN-based models to mine information from BIM log data to support BIM-based building design decisions. Zhang and El-Gohary [23] proposed a bidirectional LSTM-based method with transfer learning strategies to extract semantic and syntactic information elements from building-code requirements. Zhong et al. [24] used a bidirectional LSTM-based model with conditional random fields to extract procedural constraints from construction regulations. Amer et al. [25] used a transformer-based method to predict the relationship between look-ahead planning tasks to master-schedule activities. Li et al. [26] used hierarchical attention networks to map bridge inspection descriptions to bridge condition ratings.

2.2 Transformers and pretrained transformer-based models

A transformer is a deep learning model structure that consists of an encoder and a decoder and uses multi-head attention mechanisms [27] within the encoder or decoder (i.e., self-attention) or between them (i.e., encoder-decoder attention) to capture the dependencies between different data points. Transformer-based models consist of multiple layers of transformers to allow for learning the contextual representations of input data. Example transformer-based models include generative pretrained transformer (GPT) models (e.g., GPT-2 [28]) by OpenAI, bidirectional encoder representations from transformers (BERT) models [29] by Google and variants of BERT [e.g., a lite BERT for self-supervised learning of language representations (ALBERT) [30] and a robustly optimized BERT pretraining approach (RoBERTa) [31]], and the vision transformer (ViT) [32]. Compared to other deep learning models (e.g., RNN-based models) that were predominately used for natural language processing (NLP) tasks, transformer-based models have improved both the language modeling performance, especially in dealing with long-term dependencies in the text, and the computational efficiency in model training. These improvements result from (1) the use of multi-head attention mechanisms in the transformer layers in place of sequential model structures such as RNN [27]; and (2) the incorporation of a deep model structure (e.g., the BERT base model that consists of 12 layers of transformers and 110 million parameters [29]). Transformer-based models can be pretrained on large general-domain corpora [e.g., BooksCorpus (800M words) and English Wikipedia (2,500M words)] through unsupervised or self-supervised learning tasks, such as masked language modeling and next sentence prediction [29]. The pretrained transformer-based language models can be then finetuned on smaller, domain- or task-specific text data for downstream NLP tasks, such as sequence labeling, machine translation, and question answering (e.g., [27-29]).

Recent efforts in the construction domain have applied transformer-based models in solving problems including defect detection (e.g., [33-35]) and information extraction (e.g., [25, 36-37]). For example, Zhou et al. [35] used transformer-based models to extract features for point cloud classification to support sewer defect detection. Kim et al. [36] used transformer-based models to learn representations for extracting infrastructure damage information from textual data. However, to the best of the authors' knowledge, no efforts focused on using transformer-based models for supporting ACC.

3 State of the art and knowledge gaps in IFC-regulation semantic information alignment

The IFC schema is used to represent and share information in the AEC domain, and is the most commonly adopted format for BIM [38]. It defines an object-based information model consisting of entities, including objects (“IfcObject”), relations (“IfcRelationship”), and properties (“IfcPropertyDefinition”). To support BIM interoperability across different applications and levels of development, a model view definition (MVD), which is a selection of IFC for a specific use or workflow (e.g., [39-41]), is further established based on the overall IFC schema. However, the IFC concepts in the IFC schema or MVDs do not naturally correspond to regulatory concepts and require additional efforts for aligning or mapping the concepts, which creates a major barrier for ACC [1].

IFC-regulation semantic information alignment aims to align or link the regulatory concepts in natural language to their corresponding or related IFC concepts (e.g., IFC entities, properties, enumerated property values) by mapping or transforming one or both types of concepts. Existing research efforts for IFC-regulation semantic information alignment predominately focus on predefined rule-based or hardcoding-based methods. They can be classified into three main groups based on how the two types of information are changed during the alignment: regulation-to-IFC translation, regulation-to-IFC mapping, and IFC-to-regulation adaptation. In regulation-to-IFC translation, the building-code requirements are hardcoded into computer-processable representations that allow information representation or retrieval with the IFC schema using modeling languages such as SPARQL protocol and Resource Description Framework (RDF) query language [42], building environment rule and analysis language [43], regulatory knowledge query language [6], visual code checking language [44], and language-integrated query [45]. In regulation-to-IFC mapping, the regulatory concepts are mapped to those in the IFC schema either fully manually or using dictionaries (e.g., bSDD [4]), rules (e.g., [2, 46]), ontologies (e.g., [42, 47-48]), procedural algorithms and functions (e.g., [49]), meta-databases and applications (e.g., [50]), or black-box mechanisms (e.g., [51-53]). In IFC-to-regulation adaptation, the IFC schema or BIM file is adapted or modified to support direct alignment to building-code requirements by adding concepts from the requirements to the IFC schema [54] or by modifying existing properties in specific BIM files [55].

Despite the state-of-the-art performance achieved by the predefined rule-based and hardcoding-based IFC-regulation semantic information alignment methods, they typically require significant manual effort. Also, many of these methods lack flexibility and adaptability (e.g., due to the use of predefined mapping rules or hardcoded computer-processable requirements) and might not allow successful implementation across different MVDs, BIMs, and different types of regulatory documents (e.g., building code versus energy code). They also require updates when the IFC schema or the regulatory documents are updated [5-6]. To overcome these limitations, recent research efforts have explored the use of machine learning to facilitate IFC-regulation semantic information alignment. Many of these efforts focused on augmenting the BIM models with additional attributes and relationships for facilitating compliance checking, using classification or other approaches, to support the alignment (e.g., [9-11]). For example, Wu et al. [10] extracted invariant signatures, which uniquely define each AEC object and capture their intrinsic properties, to classify IFC objects and augment the models with the predicted/classified types. Another smaller number of efforts focused on directly aligning the regulatory concepts to the IFC concepts using machine learning approaches. For example, Zhang and El-Gohary [54] developed a semiautomated machine learning-based method to extend the IFC schema with regulatory concepts, which consists of three main steps: rule-based regulatory concept extraction, similarity-based term matching, and supervised learning-based relation classification. Zhou and El-Gohary [8] proposed a deep learning-based method for learning semantic representations of building-code and IFC concepts for information alignment of BIMs to building-code requirements, which uses semantic similarity analysis, searching, and network construction. However, the aforementioned machine learning-based approaches share three common limitations. First, despite achieving higher levels of automation and generalizability (than rule-based and hardcoding-based methods), they still require significant manual effort. For example, the semiautomated approach in [54] requires interim checking, and possibly fixing, of intermediate results by the users. Second, they mostly rely on traditional, contextless semantic representations (e.g., word embeddings such as word2vec [56] and global vectors for word representations [57]) and manually engineered features such as the part-of-speech patterns of the concepts, number of words in the concepts, and first or last term in the concepts. These features are less effective in capturing the domain-specific semantics (for example, compared to the contextual representations learned by transformer-based models), which are essential for determining the relations between concepts in semantic information alignment. Third, they do not leverage the important contextual information and knowledge contained in the IFC schema and the regulatory documents, such as the natural-language definitions of the concepts and the IFC knowledge graph, which provide additional semantic information for interpreting and aligning semantically or syntactically complex regulatory concepts.

4 Proposed transformer-based method for automated context-aware IFC-regulation semantic information alignment

A transformer-based method for automated context-aware IFC-regulation semantic information alignment for supporting ACC is proposed. First, the proposed method uses a relation classification model to align regulatory concepts extracted from building codes and standards with the concepts in the IFC schema (i.e., the IFC objects and their predefined types). The model classifies each pair of IFC-regulatory concepts as semantically related or not. For the purpose of ACC, an IFC concept is aligned/related to a regulatory concept if they are equivalent (e.g., “IfcRamp” and “ramp”) or if the IFC concept is a supertype of the regulatory concept (e.g., “IfcDoor” and “revolving door”). Aligning to superclasses is adopted for IFC-regulation alignment in ACC applications because the regulatory documents typically have more specific concept descriptions than those in the IFC. Second, the proposed method is context-aware because it (1) learns contextual representations of words using pretrained transformer-based models; and (2) leverages the natural-language definitions of the regulatory and IFC concepts and an IFC knowledge graph to provide supplemental contextual information and knowledge for finetuning pretrained transformer-based models using transfer learning.

The method is composed of five main steps, as per Fig. 1: (1) IFC knowledge graph development based on the IFC schema and the IFC ontology, (2) concept pair development based on the IFC knowledge graph, (3) transformer-based concept relation classification, (4) model training/finetuning with transfer learning strategies, and (5) post-classification concept pair pruning.

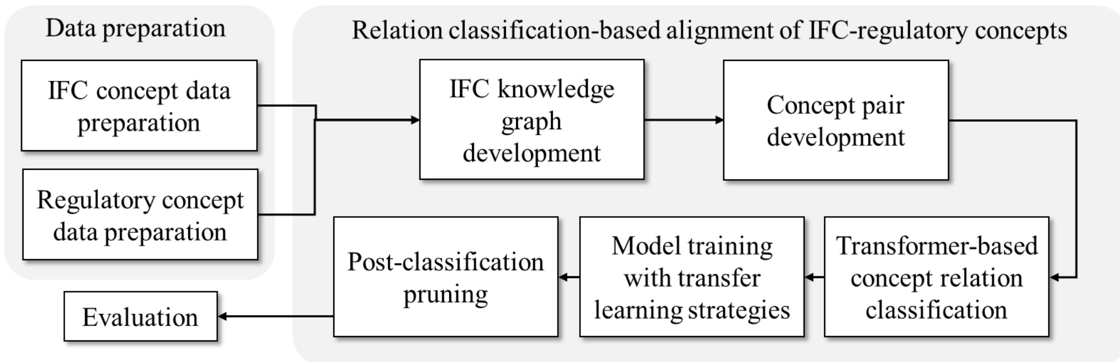


Fig. 1. Proposed transformer-based method for automated context-aware IFC-regulation semantic information alignment.

4.1 Concept data preparation

4.1.1 IFC concept data preparation

The IFC concept data were prepared to develop the concept pairs for training (for finetuning the pretrained models with domain-specific data using transfer learning) and testing the proposed method. The data were automatically prepared based on the buildingSMART International standards and supporting documentation on IFC4 using four steps: (1) collecting the .htm files of the IFC entities and property sets, (2) parsing the files, (3) extracting the natural-language canonical forms and definitions from the files, and (4) uncasing and cleaning the natural-language canonical forms and definitions of the IFC concept instances. As a result, each IFC concept data instance consists of three parts: the IFC concept name, the natural-language canonical form, and the natural-language definition. The IFC concept name is the name of the entity in the IFC schema. The natural-language canonical form is the name of the entity in a natural language (e.g., English), which is uncased and singular. The definition is the natural-language definition of the entity in the IFC schema. For example, the canonical form of "IfcDoor" is "door", and its natural-language definition is "The door is a building element that is predominately used to provide controlled access for people and goods. It includes constructions with hinged, pivoted, sliding, and additionally revolving and folding operations. A door consists of a lining and one or several panels" [38]. Table 1 shows examples of two different types of IFC concepts (i.e., entity and enumerated value) in the IFC schema version 4 and the associated data used in this study. A total of about 2,000 IFC concept instances and their data were prepared.

Table 1. Example IFC Concept Data Instances in Training and Testing Data

IFC concept	Type of IFC concept	Natural-language canonical form	Natural-language definition from IFC schema
IfcAlarm	Entity	Alarm	An alarm is a device that signals the existence of a condition or situation that is outside the boundaries of normal expectation or that activates such a device.
IfcSpatialZone	Entity	Area, space, zone	A spatial zone is a non-hierarchical and potentially overlapping decomposition of the project under some functional consideration. A spatial zone might be used to represent a thermal zone, a construction zone, a lighting zone, a usable area zone.
IfcElectricApplianceTypeEnum - REFRIGERATOR	Enumerated value	Refrigerator	An electrical appliance that has the primary function of storing food at low temperature but above the freezing point of water.
IfcDistributionSystemEnum - FIREPROTECTION	Enumerated value	Fire protection	Fire protection sprinkler system.

4.1.2 Regulatory concept data preparation

The regulatory concept data were prepared to develop the concept pairs for testing the transformer-based relation classification model. A regulatory concept data instance is defined as a sequence of words consisting of the canonical form and the definition of a regulatory concept, both of which are in the form of natural language and are directly extracted from the regulatory documents. For example, the data instance of the concept "fire-rated glazing" is the concatenation of "fire-rated glazing" and its definition "glazing with either a fire protection rating or a fire-resistance rating" [7]. The regulatory concept data were developed based on the concepts and definitions from the following chapters and sections in three different types of regulatory documents: (1) Section 202 *Definitions* of IBC, (2) Section C202 *General Definitions* and Section R202 *General Definitions* of IECC, and (3) 106.5 *Defined Terms* of ADA Standards. The natural-language canonical forms and definitions were uncased and cleaned. A total of 220 regulatory concept data instances were prepared. Table 2 shows examples of regulatory concept data from different sources [7, 58-59].

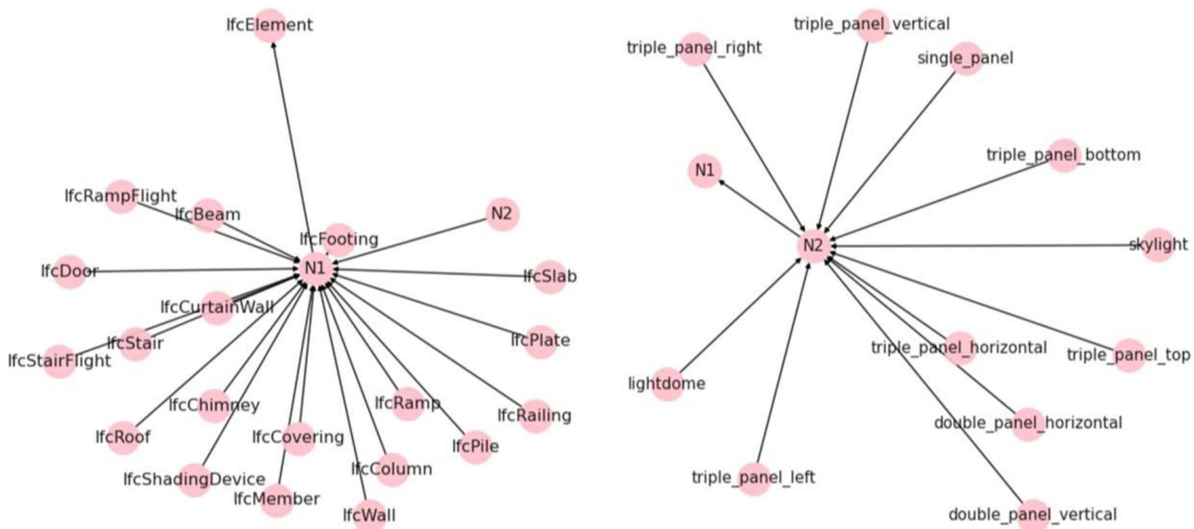
Table 2. Example Regulatory Concept Data Instances in Testing Data

Regulatory concept canonical form	Source regulatory document	Natural-language definition
Membrane-covered cable structure	International building Code (IBC)	A nonpressurized structure in which a mast and cable system provides support and tension to the membrane weather barrier and the membrane imparts stability to the structure.
Circulating hot water system	International Energy Conservation Code (IECC)	A specifically designed water distribution system where one or more pumps are operated in the service hot water piping to circulate heated water from the water-heating equipment to the fixture supply and back to the water-heating equipment.
Qualified historic building or facility	Americans with Disabilities Act Standards for Accessible Design (ADA Standards)	A building or facility that is listed in or eligible for listing in the National Register of Historic Places, or designated as historic under an appropriate State or local law.

4.2 IFC knowledge graph development

For determining the relations between the IFC concepts and accordingly developing the concept pairs (see Section 4.3), a simple IFC knowledge graph was developed based on the IFC schema and the IFC ontology [60]. The knowledge graph is a directed graph that consists of IFC concepts as nodes and the relations between pairs of concepts (e.g., "is subclass of") as edges between the nodes. Fig. 2 shows two example subgraphs induced from the IFC knowledge graph. The subgraphs consist of the neighbors that are centered at the nodes representing the IFC concepts "IfcBuildingElement" and "IfcWindow" within a radius of one.

The knowledge graph was constructed following two steps. First, a knowledge graph was automatically constructed based on the ifcOWL (Web Ontology Language representation of the ifc schema) [60], which is an RDF graph of the IFC ontology, using a rule-based method. For example, the blank nodes in the ifcOWL were removed and the edges that link the blank nodes with the uniform resource identifier (URI) reference nodes were redirected accordingly. Second, the predefined types of the IFC concepts (e.g., "triple_panel_left" as a predefined type of "IfcWindow" in Fig. 2) were added to the knowledge graph as subclasses of these IFC concepts.



Note: N1=IfcBuildingElement; N2=IfcWindow

Fig. 2. Example subgraphs centered at the IFC concepts "IfcBuildingElement" (left) and "IfcWindow" (right) induced from the IFC knowledge graph.

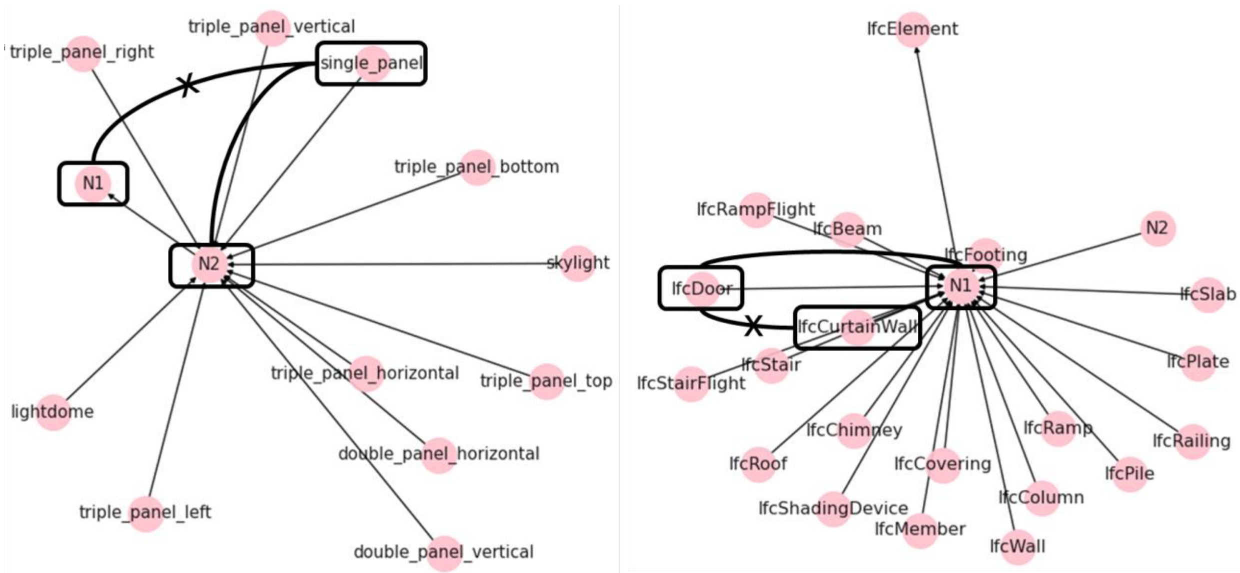
4.3 Concept pair development for training and testing

Two concept pair datasets were developed for training and testing. Fig. 3 and Table 3 show example concept pairs developed based on the IFC knowledge graph. For training, a dataset of concept pairs was developed for finetuning the pretrained model with domain-specific data using transfer learning strategies). The pairs were developed using the IFC concept data (Section 4.1.1), with the support of the developed IFC knowledge graph (Section 4.2). Each concept pair that serves as a positive training instance consists of two semantically related IFC concepts that are directly linked by one edge in the IFC knowledge graph. Each concept pair that serves as a negative training instance consists of two IFC concepts that are *not* directly linked by an edge. For example, the concept pair of the IFC concepts "IfcDoor" and "IfcBuildingElement" is related; and the concept pair of "IfcDoor" and "IfcWindow" is not related. A total of about 20,000 training concept pairs were developed.

Table 3. Example Training Concept Pairs

Concept pair (in canonical form)		Binary relation between Concepts 1 and 2
Concept 1	Concept 2	
Building element	Curtain wall	Related
Distribution control element	Flow instrument	Related
Curtain wall	Flow instrument	Not related
Building element	Distribution control element	Not related
Electric appliance	Refrigerator	Related
Refrigerator	Fire protection	Not related

For testing, a dataset of concept pairs was developed for serving as the gold standard to evaluate the proposed method. Each concept pair consists of one IFC concept and one regulatory concept, and the pairs were developed using the prepared concept data (Section 4.1). For preparing the positive testing instances, for each regulatory concept, the semantically related IFC concept(s) was manually selected by a group of three experts, one from industry and two from academia. The authors adopted a purposive sampling strategy, which aims to select a specific type of experts according to predefined criteria [61]. Two criteria were defined: (1) familiarity with building codes and compliance checking processes, and (2) familiarity with the IFC schema. The authors used purposive sampling because (1) it is suitable for small, specialized populations; and (2) it helps obtain information from a concentrated, carefully selected sample [61-62]. Each expert independently selected and paired the concepts, with an initial inter-annotator agreement of 80% in F1 measure, which indicates good consistency, reliability, and reproducibility of the process of manually aligning the regulatory and IFC concepts and thus high quality of the manual alignment for preparing the testing dataset [63-64]. The discrepancies among the annotated pairs were then resolved by the experts to reach full agreement on the final gold standard. For preparing the negative testing instances, for each regulatory concept, the IFC concepts in all ACC-relevant domains (e.g., IFC architecture domain, IFC building controls domain, and IFC structural elements domain) were enumerated and paired with the regulatory concept, except for the semantically related IFC concept(s). For example, the pair of "exit access ramp" (regulatory concept) and "IfcRamp" (IFC concept) was included as a positive instance, while the pair of "fire door" (regulatory concept) and "IfcRamp" (IFC concept) was included as a negative one. A total of 42,180 testing concept pairs, with their relations and concept definitions, were developed.



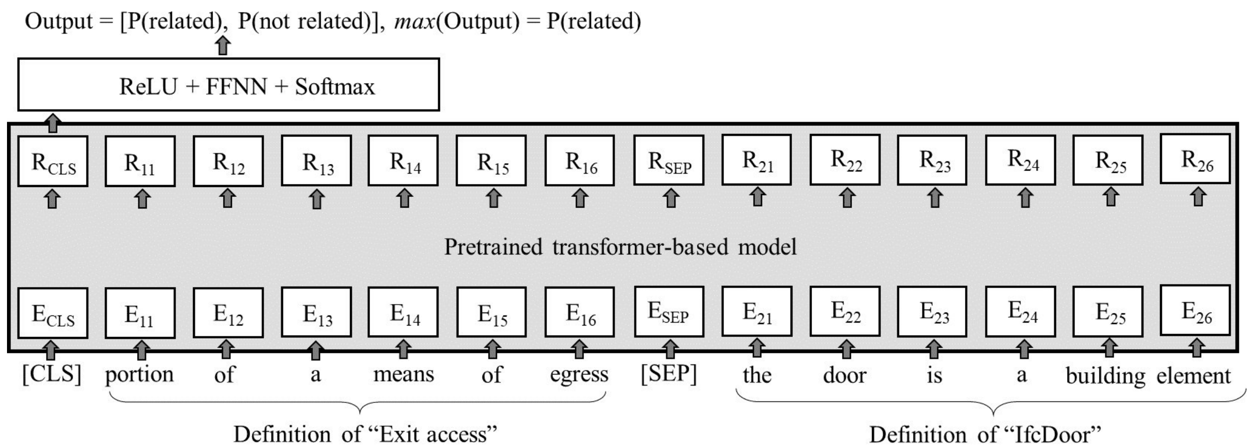
Note: N1=IfcBuildingElement; N2=IfcWindow; ▢—▢=related concept pair; ▢✕▢=not related concept pair

Fig. 3. Example related and not related concept pairs based on IFC knowledge graph.

4.4 Transformer-based concept relation classification model development

The semantic information alignment of regulatory concepts with the IFC schema is formulated as a binary relation classification problem, where given a concept pair of an IFC and a regulatory concept, a relation classification model predicts the relation between the two concepts (semantically related or not). The relation classification model consists of two main components: the pretrained transformer-based model, and a relation classification layer, which further consists of an activation function [e.g., rectified linear unit (ReLU)], a feedforward neural networks (FFNN) layer, and a softmax function, as shown in Fig. 4.

The relation classification step further consists of three substeps: definition tokenization, input sequence construction, and relation prediction. First, the natural-language definitions for the concept pairs are tokenized using the tokenizer corresponding to the pretrained transformer-based model. Second, the input to the model, which is a sequence of tokens (e.g., words and numbers), is constructed by concatenating the two tokenized definitions for each pair. The two definitions are separated by a [SEP] token, which indicates the boundary between the two definitions. The entire sequence is started with a [CLS] token, which captures the definition-level information of the relation between the two concepts through model training/finetuning with transfer learning strategies. Third, the tokens in the input sequence are embedded and loaded into the pretrained transformer-based model, which generates the output embeddings. The relation classification layer then computes the distribution over both classes, given the output embedding of the [CLS] token. The final relation predicted by the classification model is the one with the highest probability.



Note: BERT=bidirectional encoder representations from transformers; CLS=token for concept pair classification; E=input token embeddings; FFNN=feedforward neural network; R=output token embeddings; ReLU=rectified linear unit; SEP=token for separating two concepts

Fig. 4. Pretrained transformer-based concept relation classification model for IFC-regulation semantic information alignment.

4.5 Model training with transfer learning strategies

The concept relation classification model was trained (finetuning the pretrained model with domain-specific data using transfer learning strategies) to minimize the objective function – multiclass cross entropy, L, as per Eq. (1). Cross entropy describes the difference between the labels in the training data, denoted as y, and the labels predicted by the model θ, denoted as c, based on the input natural-language definitions x, as shown in Eq. (1), where D is a batch of the training data, C is the set of labels, $p_\theta(c|x_i)$ is the conditional probability of c given the input sentence x generated by the relation classification layer in the model with parameters θ, and $1_{y=c}$ is the indicator function, which returns 1 when y and c are equal, and returns 0 when y and c are not equal.

$$L(\theta) = \frac{1}{|D|} \sum_{x,y \in D} \sum_{c \in C} 1_{y=c} \log p_\theta(c|x_i) \quad (1)$$

Two transfer learning strategies to train the relation classification model were adopted for comparative evaluation: (1) the pretrained transformer-based model is not trainable, and only the relation classification layer is trainable; and (2) specific transformer layers (e.g., all the 12 layers in BERT or ALBERT base model) in the pretrained model are trainable, together with the relation classification layer. The first strategy preserves more of the semantic and syntactic information learned by the pretrained models from the general-domain text data, while the second strategy encourages learning domain- and task-specific semantic and syntactic information during the training of the model with concept pairs.

Two training practices were adopted for more stable and efficient training: (1) early stopping: the training process was stopped when the loss change is smaller than 0.1; and (2) learning rate scheduling: the learning rate was initialized small and increased as the training progresses.

4.6 Post-classification concept pair pruning

The post-classification concept pair pruning aims to select the most lexically and semantically similar IFC-regulatory concept pairs among those classified as semantically related by the relation classification model (Section 4.5) – acting like a filtering layer. The pruning consists of three main steps. First, the concept pairs were ranked according to the relation classification probabilities, which are obtained from the relation classification model. Concept pairs that are not within the top *k* of the ranking are pruned (i.e., considered not related). Second, for each classified concept pair, the word-level semantic similarity was defined as the cosine similarity between the corresponding pair of semantic concept representations of their natural-language canonical forms, as per Eq. (2), where S_c is the semantic representation of the canonical form of an IFC concept *c* and S_r is the semantic representation of the regulatory concept *r*. Concept pairs with similarities lower than a predetermined threshold (e.g., 0.9) are pruned. Third, if a regulatory concept is related to both an IFC concept and its subconcept, only the IFC subconcept is selected (to avoid redundancy, since an IFC subconcept is already related to its superconcept based on the IFC schema).

$$Similarity\ (c,r) = \frac{\mathrm{S}_c \cdot \mathrm{S}_r}{\|\mathrm{S}_c\| \|\mathrm{S}_r\|} \qquad (2)$$

4.7 *Evaluation*

For evaluating the relation classification-based semantic alignment method, three metrics were calculated for each label (semantically related or not related): precision, recall, and F1 measure, as shown in Eqs. (3) to (5), where for each label R, TP is the number of true positives (i.e., number of concept pairs correctly labeled with R), FP is the number of false positives (i.e., number of concept pairs incorrectly labeled with R), and FN is the number of false negatives (i.e., number of concept pairs not labeled with R but should have been) [65]. The overall performance of the proposed method was obtained by further calculating the average precision, recall, and F1 measure both labels.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

5 Experiments, results, and discussion

5.1 *Training and model hyperparameters*

The proposed transformer-based IFC-regulation semantic information alignment method was deployed and trained using PyTorch built in Python 3 and run using the Tesla K80 GPU provided in Google Colaboratory. A five-fold cross validation was conducted for optimizing the hyperparameters of the classification model. For the cross validation, the training data (i.e., the IFC concept pairs) were further split into two subsets – one for model training and the other for model validation. The values of other hyperparameters were determined based on the characteristics of the training and testing data used in the experiments (e.g., the maximum sentence length is 128), or the parameters of the pretrained transformer-based models (e.g., the dimension of the FFNN layer is 768 when the ALBERT base model is adopted, whose transformer layer has a dimension of 768). The values of the final training and model hyperparameters are shown in Table 4.

Table 4. Training and Model Hyperparameters for Proposed Classification Model

Hyperparameter	Value
Training	
Batch size of training data	32
Maximum length of tokenized definition pair	256
Initial learning rate	1e-5
Dropout rate	0.1
Model	
Dimension of the output layer	Same as transformer layer size (e.g., 768 for ALBERT base model)
Number of attention heads	Depending on pretrained transformer-based model (e.g., 12 for ALBERT base model)
Number of hidden layers	Depending on pretrained transformer-based model (e.g., 12 for ALBERT base model)
Hidden layer size	Depending on pretrained transformer-based model (e.g., 768 for ALBERT base model)

5.2 Application of proposed method

Fig. 5 illustrates the application of the proposed relation classification-based semantic alignment method, with an example. Given a pair of regulatory and IFC concepts and their definitions, first, the trained transformer-based concept relation classification model predicts the relation between concepts, generating candidate related regulatory and IFC concept pairs with their relation probabilities. Second, all candidate related concept pairs are ranked based on the relation probabilities. Third, given the representations of the concepts, the concept similarities are assessed by computing the cosine similarities between the representations. Fourth, the final related concept pairs are determined based on rules (e.g., the top k candidate pairs are retained as final pairs). The final related concept pairs are further used in downstream ACC tasks, such as compliance reasoning.

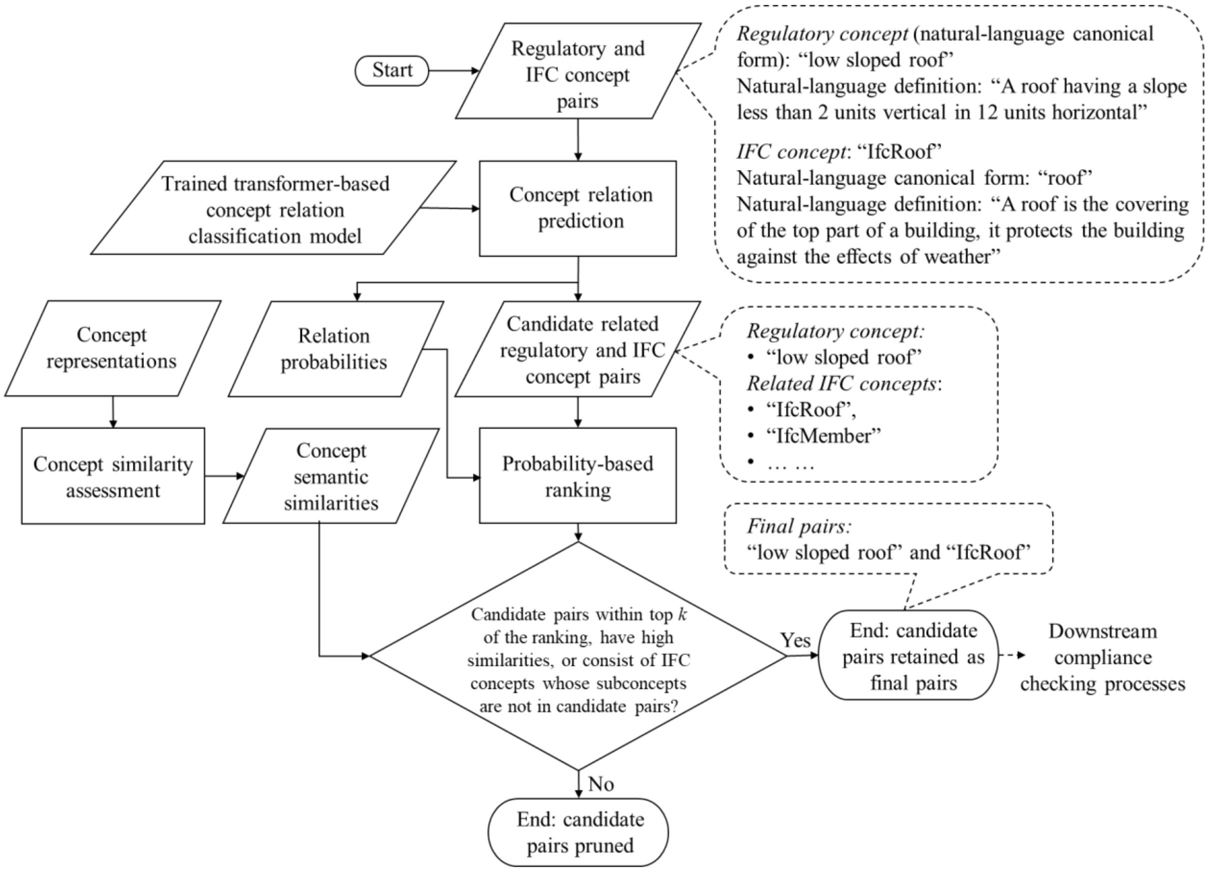
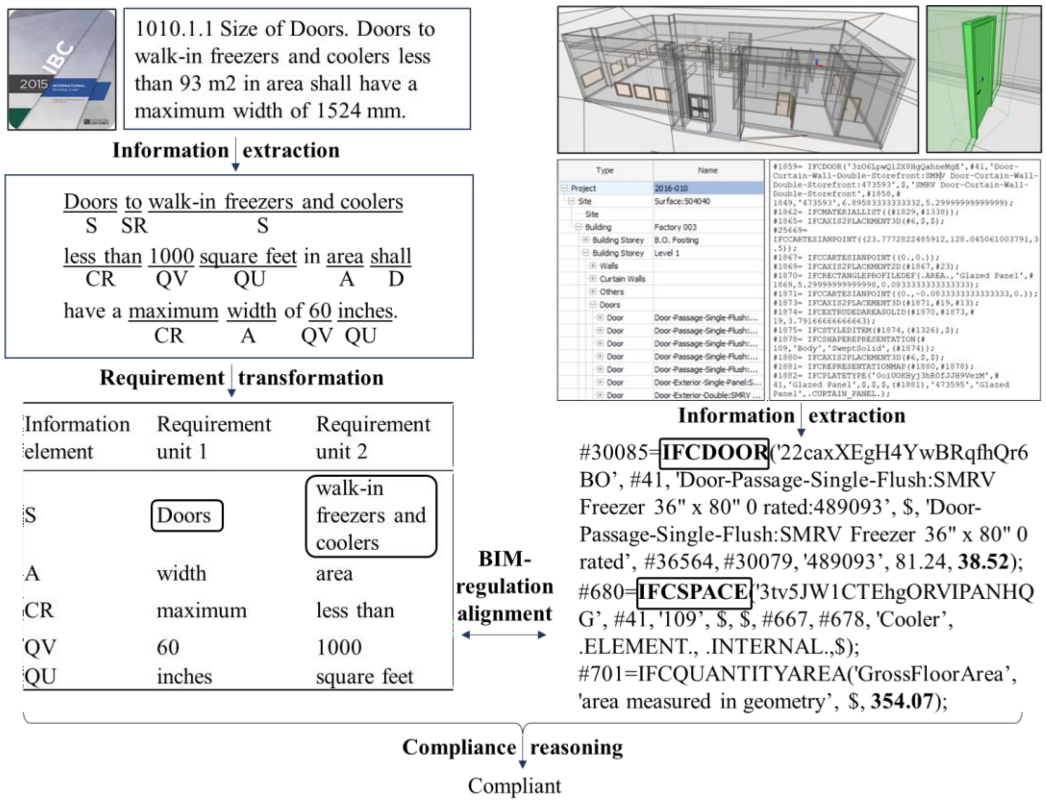


Fig. 5. Proposed semantic information alignment method.

Fig. 6 provides an example to further illustrate the use of the proposed method within an ACC system. The ACC system consists of four main modules: (1) information extraction (regulatory information [23] and design/BIM information [66]), (2) requirement transformation [67], (3) BIM-regulation alignment, and (4) compliance reasoning [66]. The proposed method can be used within the BIM-regulation alignment module to align the regulatory concepts in the extracted and transformed requirements (output of module 2) to the IFC concepts in the IFC instances (output of module 1). The aligned requirements and IFC instances (output of module 3) are the input to the final rule-based compliance reasoning module (module 4), where the information (e.g., compliance checking attributes such as area and width) in the requirements are compared to the information in the IFC instances to determine the compliance results. For the details of modules 1, 2, and 4, the readers are referred to [23, 66-67].



Note: S=Subject; SR=Subject relation; CR=Comparative relation; QV=Quantity value; QU=quantity unit; A=Compliance checking attribute; D=Deontic operator indicator

Fig. 6. Example to illustrate use of proposed method for BIM-regulation alignment within an automated compliance checking (ACC) system.

5.3 Evaluation of information alignment performance

The testing data (see Section 4.3) were used to evaluate the performance of the proposed method. Four sets of ablation experiments (Sections 5.3.1 to 5.3.4) were conducted to better understand the impact of four important aspects on the performance of the proposed method: (1) the different types of pretrained transformer-based models, (2) the process of training/finetuning the relation classification model using transfer learning strategies, (3) the incorporation of natural-language definitions as contextual information for training the classification model, and (4) the post-classification concept pair pruning. A fifth set of experiments (Section 5.3.5) was conducted to assess the performance of the proposed method across different types of regulatory documents. The final selected model uses the ALBERT base pretrained model with 12 trainable transformer layers, natural-language definitions of IFC and regulatory concepts, and a threshold of 5 for top-k in post-classification pruning. It achieved average precision, recall, and F1 measure of 84.3%, 83.3%, and 83.8%, respectively.

5.3.1 Impact of different types of pretrained transformer-based models

The proposed method was tested with different types of pretrained transformer-based models (i.e., BERT and ALBERT) and models of different sizes. Four different pretrained transformer-based models were tested: ALBERT base (12 transformer layers, 768-layer size, and 11 million parameters), ALBERT large (24 transformer layers, 1024-layer size, and 17 million parameters), ALBERT xlarge (24 transformer layers, 2048-layer size, and 58 million parameters), and BERT base (12 transformer layers, 768-layer size, and 110 million parameters) models.

As shown in Table 5, the proposed method with the ALBERT base model performed the best in terms of average precision, recall, and F1 measure, outperforming the proposed method with other pretrained models, by an average of 14.4% in precision, 20.8% in recall, and 18.5% in F1 measure. The experimental results indicate that for the specific training data used and the specific relation prediction task, the ALBERT base model is of the most suitable size, while larger models might start to overfit or underfit. A large model (i.e., the ALBERT large model) achieved lower performance, especially lower recall, compared to the base model, and thus was not selected because few false negatives and a high recall are required for ACC tasks.

Table 5. Performance of Proposed Method with Different Pretrained Transformer-based Models

Pretrained transformer-based models	Precision	Recall	F1 measure
ALBERT base model	**84.3%**	**83.3%**	**83.8%**
ALBERT large model	81.5%	70.2%	74.6%
ALBERT xlarge model	76.7%	65.7%	69.8%
BERT base model	51.5%	51.5%	51.5%

Note: Bolded font indicates highest performance; 12 trainable transformer layers, natural-language definitions of IFC and regulatory concepts, and a threshold of 5 for top-k in post-classification pruning were used.

5.3.2 Impact of different transfer learning strategies for pretrained transformer-based relation classification

The proposed method was tested with different transfer learning strategies for training/finetuning the pretrained transformer-based relation classification model for assessing the impact of balancing domain-general and domain-specific semantic and syntactic information on performance. Two different transfer learning strategies were tested: fixing or training the pretrained transformer-based model in the relation classification model. For the second strategy, different numbers of trainable transformer layers were also tested for comparative evaluation. The ALBERT base model was used in this set of experiments.

As shown in Table 6, the proposed method with the trainable pretrained transformer-based model, and with twelve trainable transformer layers, showed the best performance in terms of average precision, recall, and F1 measure, outperforming the proposed method when the other strategies were adopted, by an average of 12.8% in precision, 18.2% in recall, and 16.5% in F1 measure. The experimental results indicate that the general-domain semantic and syntactic information transferred by the pretrained models is not sufficient for relation classification with complex regulatory concepts, and that part of the pretrained models (e.g., the last transformer layers) need to be trainable to adapt itself to domain- and task-specific data. The model with less trainable layers achieved lower performance, especially lower recall, compared to the one with 12 trainable layers. The latter model was, thus, selected because of the higher priority need for recall. The experimental results also indicate that the representations learned through training/finetuning pretrained transformer-based models could serve as an important source of contextual information that could contribute to an increase of around 30.0% in relation classification performance (in terms of precision, recall, and F1 measure).

Table 6. Performance of Proposed Method with Different Finetuning Strategies with Pretrained Transformer-based Models

Transfer learning strategies for training the relation classification model	Number of trainable transformer layers	Precision	Recall	F1 measure
Fixed pretrained transformer-based model	0	58.7%	52.0%	53.2%
	4	77.7%	73.3%	75.3%
Trainable pretrained transformer-based model	8	78.0%	70.0%	73.3%
	12	**84.3%**	**83.3%**	**83.8%**

Note: Bolded font indicates highest performance; the pretrained ALBERT base model, natural-language definitions of IFC and regulatory concepts, and a threshold of 5 for top-*k* in post-classification pruning were used.

5.3.3 Impact of contextual text data

The proposed method was tested with different IFC and regulatory concept data to assess the impact of utilizing the natural-language definitions in the proposed method. Four different types of data were tested: (1) only canonical forms for both IFC and regulatory concepts, (2) canonical forms and definitions for both IFC and regulatory concepts (the proposed types of concept data), (3) canonical forms and definitions for regulatory concepts, and only canonical forms for IFC concepts, and (4) canonical forms and definitions for IFC concepts, and only canonical forms for regulatory concepts.

As shown in Table 7, the proposed method with the proposed form of concept data (i.e., concept data with both natural-language canonical forms and definitions for both IFC and regulatory concepts) showed the best performance in terms of average precision, recall, and F1 measure, outperforming the proposed method when other types of concept data were used, by an average of 29.5% in precision, 29.6% in recall, and 29.9% in F1 measure. The experimental results indicate that the definitions could serve as an important source of contextual information that could be captured and leveraged by the transformer-based models through transfer learning and could contribute to an increase of over 30.0% in relation classification performance.

Table 7. Performance of Proposed Method with Different Types of Concept Data

Contextual information included in concept data	Precision	Recall	F1 measure
Natural-language canonical forms for IFC and regulatory concepts	53.3%	50.8%	51.3%
Natural-language canonical forms and definitions for IFC and regulatory concepts	**84.3%**	**83.3%**	**83.8%**
Natural-language canonical forms and definitions for IFC concepts and only natural-language canonical forms for regulatory concepts	60.2%	60.2%	60.2%
Only natural-language canonical forms for IFC concepts and natural-language canonical forms and definitions for regulatory concepts	50.9%	50.2%	50.2%

Note: Bolded font indicates highest performance; the pretrained ALBERT base model with 12 trainable transformer layers and a threshold of 5 for top-*k* in post-classification pruning were used.

5.3.4 Impact of post-classification pruning

The proposed method was tested with different post-classification pruning thresholds for assessing the impact of pruning on performance. Five different thresholds for top-*k* pruning using both the relation classification probability-based ranking and the word-level semantic similarity-based ranking were tested: one, three, five, seven, and nine.

As shown in Table 8, the proposed method with a threshold of 5 for top-*k* pruning showed the best performance in terms of average precision, recall, and F1 measure, outperforming the proposed method with other thresholds, by an average of 5.4% in precision, 4.8% in recall, and 5.1% in F1 measure. The experimental results indicate that a threshold of 5 was optimal in this case, because it retained more true positives compared to smaller thresholds and excluded more false positives compared to larger thresholds.

Table 8. Performance of Proposed Method with Different Post-classification Concept Pair Pruning Thresholds

Threshold for top-*k* pruning	Precision	Recall	F1 measure
1	78.0%	77.6%	77.8%
3	80.0%	79.6%	79.8%
5	**84.3%**	**83.3%**	**83.8%**
7	79.1%	78.7%	78.9%
9	78.4%	78.0%	78.2%

Note: Bolded font indicates highest performance; the pretrained ALBERT base model with 12 trainable transformer layers and natural-language definitions of IFC and regulatory concepts were used.

5.3.5 Performance of the proposed method across different types of documents

The proposed method was tested on regulatory concepts extracted from three different types of documents for assessing its performance across different codes and standards: IBC, IECC, and ADA Standards. As shown in Table 9, the proposed method achieved good performance across the three documents, in terms of average precision, recall, and F1 measure. A relatively lower performance (about 8-9% in F1 measure) was shown for IBC and IECC, compared to ADA Standards, which is likely due to the relatively high complexity (e.g., complex noun phrases and verb phrases) of some of the regulatory concepts contained in the two documents.

Table 9. Performance of Proposed Method on Different Types of Regulatory Documents

Type of regulatory document	Precision	Recall	F1 measure
International Building code (IBC)	82.7%	81.3%	81.9%
International Energy Conservation Code (IECC)	82.5%	82.5%	82.5%
Americans with Disabilities Act Standards (ADA Standards)	91.4%	90.4%	90.9%

Note: The pretrained ALBERT base model with 12 trainable transformer layers, natural-language definitions of IFC and regulatory concepts, and a threshold of 5 for top-*k* in post-classification pruning were used.

5.4 Error Analysis

Three main sources of errors were identified based on the experimental results. First, the proposed method had errors when dealing with regulatory concepts whose corresponding canonical forms are less frequent in the regulatory document, such as “sallyport”, which appears less than ten times in only one section of the IBC. The low performance is likely because the transformer-based models were pretrained on general-domain text data where such words rarely appear and thus the models are less capable to capture their semantic information. Second, the proposed method showed relatively lower performance for regulatory concepts that have definitions that are semantically or syntactically very complex (e.g., long, complex definition with multiple or recursive conditions) or very simple (e.g., simple definition consisting of only a few words). The lower performance is due to the high syntactic complexity (e.g., complex noun phrases, verb phrases, and preposition phrases, and clauses of different types) and high semantic complexity (e.g., having multiple references and restrictions) of the complex definitions, or the lack of sufficient semantic information provided in the simple definitions. Third, the proposed method showed relatively lower performance for concepts from IBC and IECC compared to those from the ADA Standards. The lower performance is due to (1) the relatively low lexical and semantic similarity between the IBC and IECC concept data and the training data developed based on the IFC knowledge graph; and (2) the relatively high complexity (e.g., complex noun phrases and verb phrases) of some of the IBC and IECC concepts.

5.5 Limitations

Three limitations of the work are acknowledged. First, the proposed method successfully leveraged contextual information, including concept definitions and existing relations between IFC concepts, for improved information alignment; however, it did not consider cases where concepts might have different definitions/meanings across different regulations or subdomains of knowledge. Additional evaluation efforts are needed to test the proposed method on other types of regulatory documents (e.g., International Fire Code) and domains (e.g., fire safety). The experimental results are expected to show similar performance; however, the performance level may vary due to possible differences in the syntactic and semantic characteristics of the concepts in those documents or domains. Second, the proposed method was tested on IFC and regulatory concepts with natural-language definitions but not on those without explicit definitions. Future efforts are needed to deal with concepts that lack such explicit definitions. This could be possibly through integrating additional external knowledge as contextual information, such as ontological and relational knowledge from other types of classification systems (e.g., Uniclass and Omniclass), natural-language descriptions or definitions of concepts from data dictionaries, encyclopedias, and specifications (e.g., bsDD). Third, the scope of the work was limited to IFC objects (e.g., IfcBuildingElement, IfcDistributionElement, IfcSpace). In future work, the proposed method could be extended to include the attributes and properties of the IFC objects (e.g., OverallHeight and OverallWidth for IfcDoor) and the IFC relations (e.g., IfcRelAggregates, IfcRelContained, IfcRelVoidsElement). For attributes and properties, a similar transformer-based context-aware approach could be used, although additional external knowledge may be needed (as contextual information) because many of the attributes and properties lack explicit natural-language definitions. For relation alignment, given the large difference in the representation/terminology of relations across the natural-language text and the IFC schema, more advanced machine learning and/or network modeling approaches could be explored.

6 Contribution to the body of knowledge

This paper offers a new method for IFC-regulation semantic information alignment. The proposed method uses a relation classification model to relate and align the IFC and regulatory concepts, which utilizes deep learning and transfer learning techniques. The proposed method showed good performance across regulatory concepts from different types of codes and standards, including IBC, IECC, and ADA Standards. The proposed method contributes to the body of knowledge in four main ways. First, it is the first effort to use pretrained transformer-based models in text and knowledge analytics for supporting ACC. It leverages these models in both predicting relations between concepts and generating concept semantic similarities for pruning candidate concept pairs. These models are able to learn contextual representations that have superior ability in capturing semantic and syntactic dependencies from text data compared to traditional contextless and/or manually engineered features. Second, the research makes use of both general-domain and domain-specific semantic and syntactic information by training/finetuning the relation classification model with transfer learning strategies. Incorporating both types of information enhances the relation classification performance and increases the scalability and flexibility of the model. Third, it innovatively leverages the natural-language definitions of the concepts for information alignment of IFC and regulatory concepts. The definitions provide contextual lexical, syntactic, and semantic information for improved relation classification and thus improved information alignment. Fourth, it also leverages the IFC knowledge graph to develop training concept pairs, which incorporates the ontological contextual knowledge. The use of knowledge graph not only reduces the manual effort in preparing the training data and thus facilitates the automation of the information alignment process, but also enables leveraging the knowledge within the IFC schema to link the IFC-regulation concept pairs for improved relation classification and thus improved information alignment.

7 Conclusions and future work

In this paper, a transformer-based method for automated context-aware IFC-regulation semantic information alignment was proposed. The proposed method uses a relation classification model to relate and align the regulatory concepts extracted from building codes and standards with the concepts in the IFC schema, where the natural-language definitions of the two sets of concepts and an IFC knowledge graph are used to provide supplemental contextual information and knowledge for finetuning a pretrained transformer-based model using transfer learning. The relation classification model was trained on IFC concept pairs consisting of natural-language canonical forms and definitions that were constructed automatically based on an IFC knowledge graph. The proposed method was tested using a developed gold-standard dataset that consists of 42,180 IFC-regulatory concept pairs. An average precision of 84.3%, recall of 83.3%, and F1 measure of 83.8% in alignment was achieved.

The analysis of the experimental results indicates that (1) it is important to adapt existing pretrained transformer-based models using domain- and task-specific data to capture the semantic and syntactic information that is specific to the data at hand for improved performance; (2) the natural-language definitions and the IFC knowledge graph provided important sources of contextual information that could be leveraged by the transformer-based models for improved classification; and (3) the proposed relation classification method showed good performance across different types of regulatory documents (IBC, IECC, and ADA Standards).

In the future, the authors plan to focus on improving the proposed method in four directions. First, the relation classification could be improved by (1) injecting more contextual information or knowledge by refining the IFC knowledge graph and incorporating more concept definitions; (2) creating more training concept pairs from both IFC schema and other resources such as bSDD; and (3) increase the scale and diversity of the testing IFC-regulatory concept pairs. Such improvements could greatly increase the model's ability to deal with complex or rare concepts. Second, the post-classification pruning could be improved by (1) incorporating additional types of representations for computing word representations, such as the representations generated by transformer layers other than the final layer; (2) exploring different weighting strategies for computing concept representations based on word representations; and (3) exploring different ranking strategies for pruning. This could help better leverage the semantic information learned by the pretrained transformer-based models with general-domain text data. Third, the information alignment process could be improved by exploring other more fine-grained classification systems, such as Omniclass and Uniclass, to facilitate bridging the gap between the natural-language regulatory concepts and the computer-processable building designs. Fourth, and most importantly, the authors plan to integrate the proposed method with other ACC methods, such as methods for regulatory text analytics (e.g., regulatory text classification, information extraction, and transformation), BIM information analytics, and compliance reasoning, in an integrated ACC platform. The planned ACC platform will consist of four modules to: (1) fully automatically process, interpret, and understand building-code requirements that are in the form of natural language, (2) transform the requirements into computer-processable forms, (3) align the representations of the requirements with the representations of the IFC-based building designs (using the proposed method), and (4) perform compliance reasoning to determine whether the building designs comply with the requirements. Our ultimate goal is to leverage deep learning, text and knowledge analytics, and other artificial intelligence approaches to reach a level where we can fully automatically process, represent, and understand the entire regulatory documents in the AEC domain and align and integrate them with the BIM-based designs for fully ACC.

8 Acknowledgements

The authors would like to thank the National Science Foundation (NSF). This material is based on work supported by the NSF under Grant No. 1827733. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

9 Data availability statement

The data generated and used during the study are available from the following link: https://publish.illinois.edu/rzhang65-data-sharing/

10 References

[1] Eastman, C., Lee, J.M., Jeong, Y.S. and Lee, J.K., 2009. Automatic rule-based checking of building designs. Automation in construction, 18(8), pp.1011-1033. https://doi.org/10.1016/j.autcon.2009.07.002

[2] Pauwels, P., Van Deursen, D., Verstraeten, R., De Roo, J., De Meyer, R., Van de Walle, R. and Van Campenhout, J., 2011. A semantic rule checking environment for building performance checking. Automation in construction, 20(5), pp.506-518. https://doi.org/10.1016/j.autcon.2010.11.017

[3] Sacks, R., Girolami, M. and Brilakis, I., 2020. Building Information Modelling, Artificial Intelligence and Construction Tech. Developments in the Built Environment, pp.100011. https://doi.org/10.1016/j.dibe.2020.100011

[4] buildingSMART, 2021a. buildingSMART Data Dictionary. http://bsdd.buildingsmart.org/#peregrine/about. (July 15, 2021).

[5] Garrett Jr, J.H., Palmer, M.E. and Demir, S., 2014. Delivering the infrastructure for digital building regulations. Journal of Computing in Civil Engineering, 28(2), pp.167-169. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000369

[6] Dimyadi, J., Pauwels, P. and Amor, R., 2016. Modelling and accessing regulatory knowledge for computer-assisted compliance audit. Journal of information technology in construction, 21, pp.317-336. http://hdl.handle.net/1854/LU-8041842

[7] ICC (International Code Council). 2018a. 2018 International Building Code. ICC. Washington, D.C. ISBN: 978-1-60983-735-8

[8] Zhou, P. and El-Gohary, N., 2021. Semantic information alignment of BIMs to computer-interpretable regulations using ontologies and deep learning. Advanced Engineering Informatics, 48, pp.101239. https://doi.org/10.1016/j.aei.2020.101239

[9] Gao, H., Zhong, B., Luo, H. and Chen, W., 2022. Computational Geometric Approach for BIM Semantic Enrichment to Support Automated Underground Garage Compliance Checking. Journal of Construction Engineering and Management, 148(1), pp.05021013. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002230

[10] Wu, J., Akanbi, T. and Zhang, J., 2022. Constructing Invariant Signatures for AEC Objects to Support BIM-Based Analysis Automation through Object Classification. Journal of Computing in Civil Engineering, 36(4), pp.04022008. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001012

[11] Wang, Z., Sacks, R. and Yeung, T., 2022. Exploring graph neural networks for semantic enrichment: Room type classification. Automation in Construction, 134, pp.104039. https://doi.org/10.1016/j.autcon.2021.104039

[12] LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature. 521(7553), pp. 436. https://doi.org/10.1038/nature14539

[13] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C., 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[14] Clark, K., Luong, M.T., Manning, C.D. and Le, Q.V., 2018. Semi-supervised sequence modeling with cross-view training. arXiv preprint arXiv:1809.08370.

[15] Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pp. 3104-3112. arXiv:1409.3215

[16] Zhang, N., Deng, S., Sun, Z., Chen, J., Zhang, W. and Chen, H., 2020, April. Relation adversarial network for low resource knowledge graph completion. In Proceedings of The Web Conference 2020, pp. 1-12. https://doi.org/10.1145/3366423.3380089

[17] Li, L., Gan, Z., Cheng, Y. and Liu, J., 2019. Relation-aware graph attention network for visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10313-10322. arXiv:1903.12314

[18] Yang, Z., Ding, M., Zhou, C., Yang, H., Zhou, J. and Tang, J., 2020, August. Understanding negative sampling in graph representation learning. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1666-1676. https://doi.org/10.1145/3394486.3403218

[19] Hassani, K. and Khasahmadi, A.H., 2020, November. Contrastive multi-view representation learning on graphs. In International Conference on Machine Learning, pp. 4116-4126. arXiv:2006.05582

[20] Huang, X., Zhang, J., Li, D. and Li, P., 2019, January. Knowledge graph embedding based question answering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 105-113. https://doi.org/10.1145/3289600.3290956

[21] Wang, X., He, X., Cao, Y., Liu, M. and Chua, T.S., 2019, July. Kgat: Knowledge graph attention network for recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 950-958. https://doi.org/10.1145/3292500.3330989

[22] Pan, Y. and Zhang, L., 2020. BIM log mining: Learning and predicting design commands. Automation in Construction, 112, pp.103107. https://doi.org/10.1016/j.autcon.2020.103107

[23] Zhang, R. and El-Gohary, N., 2021. A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. Automation in Construction, 132, pp.103834. https://doi.org/10.1016/j.autcon.2021.103834

[24] Zhong, B., Xing, X., Luo, H., Zhou, Q., Li, H., Rose, T. and Fang, W., 2020. Deep learning-based extraction of construction procedural constraints from construction regulations. Advanced Engineering Informatics, 43, pp.101003. https://doi.org/10.1016/j.aei.2019.101003

[25] Amer, F., Jung, Y. and Golparvar-Fard, M., 2021. Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction. Automation in Construction, 132, pp.103929. https://doi.org/10.1016/j.autcon.2021.103929

[26] Li, T., Alipour, M. and Harris, D.K., 2021. Mapping textual descriptions to condition ratings to assist bridge inspection and condition assessment using hierarchical attention. Automation in Construction, 129, pp.103801. https://doi.org/10.1016/j.autcon.2021.103801

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008. arXiv:1706.03762

[28] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8), pp.9. http://www.persagen.com/files/misc/radford2019language.pdf

[29] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[30] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[31] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[32] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[33] Wang, W. and Su, C., 2022. Automatic concrete crack segmentation model based on transformer. Automation in Construction, 139, pp.104275. https://doi.org/10.1016/j.autcon.2022.104275

[34] Shamsabadi, E.A., Xu, C., Rao, A.S., Nguyen, T., Ngo, T. and Dias-da-Costa, D., 2022. Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. Automation in Construction, 140, pp.104316. https://doi.org/10.1016/j.autcon.2022.104316

[35] Zhou, Y., Ji, A. and Zhang, L., 2022. Sewer defect detection from 3D point clouds using a transformer-based deep learning model. Automation in Construction, 136, pp.104163. https://doi.org/10.1016/j.autcon.2022.104163

[36] Kim, Y., Bang, S., Sohn, J. and Kim, H., 2022. Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers. Automation in Construction, 134, pp.104061. https://doi.org/10.1016/j.autcon.2021.104061

[37] Wu, H., Shen, G.Q., Lin, X., Li, M. and Li, C.Z., 2021. A transformer-based deep learning model for recognizing communication-oriented entities from patents of ICT in construction. Automation in Construction, 125, pp.103608. https://doi.org/10.1016/j.autcon.2021.103608

[38] buildingSMART, 2021b. Industry Foundation Classes. https://standards.buildingsmart.org/IFC/DEV/IFC4_2/FINAL/HTML/. (July 15, 2021).

[39] Gui, N., Wang, C., Qiu, Z., Gui, W. and Deconinck, G., 2019. IFC-based partial data model retrieval for distributed collaborative design. Journal of Computing in Civil Engineering, 33(3), pp.04019016. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000829

[40] Akanbi, T., Zhang, J. and Lee, Y.C., 2020. Data-Driven Reverse Engineering Algorithm Development Method for Developing Interoperable Quantity Takeoff Algorithms Using IFC-Based BIM. Journal of Computing in Civil Engineering, 34(5), pp.04020036. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000909

[41] Lee, Y.C., Shariatfar, M., Ghannad, P., Zhang, J. and Lee, J.K., 2020. Generation of Entity-Based Integrated Model View Definition Modules for the Development of New BIM Data Exchange Standards. Journal of Computing in Civil Engineering, 34(3), pp.04020011. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000888

[42] Yurchyshyna, A. and Zarli, A., 2009. An ontology-based approach for formalisation and semantic organisation of conformance requirements in construction. Automation in Construction, 18(8), pp.1084-1098. https://doi.org/10.1016/j.autcon.2009.07.008

[43] Lee, Y.C., Eastman, C.M. and Lee, J.K., 2015. Automated rule-based checking for the validation of accessibility and visibility of a building information model. In Computing in Civil Engineering 2015, pp. 572-579. https://doi.org/10.1061/9780784479247.071

[44] Preidel, C. and Borrmann, A., 2016. Towards code compliance checking on the basis of a visual programming language. Journal of Information Technology in Construction (ITcon), 21(25), pp.402-421. https://www.itcon.org/papers/2016_25.content.01707.pdf

[45] Nawari, N.O., 2020. Generalized adaptive framework for computerizing the building design review process. Journal of Architectural Engineering, 26(1), pp.04019026. https://doi.org/10.1061/(ASCE)AE.1943-5568.0000382

[46] Tan, X., Hammad, A. and Fazio, P., 2010. Automated code compliance checking for building envelope design. Journal of Computing in Civil Engineering, 24(2), pp.203-211. https://doi.org/10.1061/(ASCE)0887-3801(2010)24:2(203)

[47] Zhong, B.T., Ding, L.Y., Love, P.E. and Luo, H.B., 2015. An ontological approach for technical plan definition and verification in construction. Automation in Construction, 55, pp.47-57. https://doi.org/10.1016/j.autcon.2015.02.002

[48] Beach, T.H., Rezgui, Y., Li, H. and Kasim, T., 2015. A rule-based semantic approach for automated regulatory compliance in the construction sector. Expert Systems with Applications, 42(12), pp.5219-5231. https://doi.org/10.1016/j.eswa.2015.02.029

[49] Delis, E.A. and Delis, A., 1995. Automatic fire-code checking using expert-system technology. Journal of computing in civil engineering, 9(2), pp.141-156. https://doi.org/10.1061/(ASCE)0887-3801(1995)9:2(141)

[50] Lee, H., Lee, J.K., Park, S. and Kim, I., 2016. Translating building legislation into a computer-executable format for evaluating building permit requirements. Automation in Construction, 71, pp.49-61. https://doi.org/10.1016/j.autcon.2016.04.008

[51] Hjelseth, E. and Nisbet, N., 2011. Exploring semantic based model checking. http://itc.scix.net/data/works/att/w78-2010-54.pdf (April 15, 2022).

[52] Solibri, 2021. Solibri Office. https://www.solibri.com/solibri-office (April 15, 2022).

[53] SMARTreview, 2021. SMARTreview. https://smartreview.biz/home (April 15, 2022).

[54] Zhang, J. and El-Gohary, N.M., 2016. Extending building information models semiautomatically using semantic natural language processing techniques. Journal of Computing in Civil Engineering, 30(5), pp.C4016004. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000536

[55] Choi, J., Choi, J. and Kim, I., 2014. Development of BIM-based evacuation regulation checking system for high-rise and complex buildings. Automation in Construction, 46, pp.38-49. https://doi.org/10.1016/j.autcon.2013.12.005

[56] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119. arXiv:1310.4546

[57] Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 10.3115/v1/D14-1162

[58] ICC (International Code Council). 2018b. 2018 International Energy Conservation Code. ICC. Washington, D.C. ISBN: 978-1-60983-749-5

[59] DOJ (U.S. Department of Justice). 2010. Americans with Disabilities Act Standards for Accessible Design. https://www.ada.gov/regs2010/2010ADAStandards/Guidance2010ADAstandards.htm. (July 15, 2021).

[60] Pauwels, P. and Terkaj, W., 2016. EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology. Automation in construction, 63, pp.100-133. https://doi.org/10.1016/j.autcon.2015.12.003

[61] Clark, V., and Creswell, J., 2008. The mixed methods readers, Sage Publications, Thousand Oaks. ISBN: 9781412951456

[62] Etikan, I., Musa, S.A. and Alkassim, R.S., 2016. Comparison of convenience sampling and purposive sampling. American journal of theoretical and applied statistics, 5(1), pp.1-4. 10.11648/j.ajtas.20160501.11

[63] Stemler, S.E., 2004. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation, 9(1), pp.4. https://doi.org/10.7275/96jp-xz07

[64] J.P. Pestian, L. Deleger, G.K. Savova, J.W. Dexheimer, I. Solti Natural language processing—the basics Pediatric Biomedical Informatics: Computer Applications in Pediatric Research, Springer,

Netherlands, Dordrecht (2012), pp. 149-172, 10.1007/978-94-007-5149-1_9 ISBN 978-94-007-5149-1

[65] Zhai, C. and Massung, S., 2016. Text data management and analysis: a practical introduction to information retrieval and text mining. Morgan & Claypool. ISBN: 9781970001167

[66] Zhang, J. and El-Gohary, N.M., 2017. Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking. Automation in construction, 73, pp.45-57. https://doi.org/10.1016/j.autcon.2016.08.027

[67] Zhang, R., and El-Gohary, N., 2022. Hierarchical representation and deep learning-based method for automatically transforming textual building codes into semantic computable requirements. Journal of Computing in Civil Engineering, in press.