Semantic Representation Learning and Information Integration of BIM and Regulations

Ruichuan Zhang, M.S., S.M.ASCE[1]
Nora El-Gohary, Ph.D., A.M.ASCE[2]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Avenue, Urbana, IL 61801; e-mail: rzhang65@illinois.edu
[2] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Avenue, Urbana, IL 61801; e-mail: gohary@illinois.edu

ABSTRACT

Automated checking of the compliance of building information modeling (BIM)-based building designs with relevant codes and regulations requires bridging the semantic gap between the Industry Foundation Classes (IFC) schema and the natural language. In most of the existing automated compliance checking (ACC) systems, the integration of the IFC schema and natural language is realized through hardcoding or predefined rules, ontologies, or dictionaries. These methods require intensive manual engineering effort and are often rigid and difficult to generalize. There is, thus, a need for an automated, and meanwhile flexible and generalizable information integration method. To address this need, this paper leverages transformer-based language models to learn the semantic representations of concepts in the building information models (BIMs) and regulatory documents. An automated IFC-regulatory information integration approach based on these learned semantic representations is proposed. The preliminary experimental results show that the proposed approach achieved promising performance – an accuracy of 80% – on integrating IFC and regulatory concepts.

INTRODUCTION

BIMs and regulatory documents such as building codes, specifications, and standards speak two different languages – the language of the IFC schema and the natural language. Thus, checking the compliance of BIM-based building designs with relevant regulations first requires bridging the semantic gap by integrating the semantic information in the natural-language requirements with that in the BIMs (Zhang and El-Gohary 2016). In most of the existing ACC systems, such information integration is realized through hardcoding (e.g., using modeling languages) methods or methods based on predefined rules, ontologies, or dictionaries. For example, the buildingSMART Data Dictionary (bSDD) (buildingSMART 2020a), an online data dictionary that contains objects and their properties for the building and construction industry, was developed to facilitate the alignment of concepts in natural-language requirements to their corresponding IFC concepts (e.g., IFC entities, properties, or enumerated property values, etc.). These methods require intensive manual engineering effort and are by nature rigid and difficult to generalize (Zhou and El-Gohary 2020). For example, the rules, ontologies, or dictionaries designed based on one chapter or document might need to be modified when applied to a different one. There is, thus, a need for an automated, and meanwhile flexible and generalizable information integration method, for fully automated compliance checking.

To address this need, machine learning-based methods have been developed and applied to solving the BIM-regulatory information integration problem, such as IFC and regulatory concept alignment (e.g., Zhou and El-Gohary 2020), IFC schema extension (e.g., Zhang and El-Gohary 2017), and IFC semantic enrichment (e.g., Wu and Zhang 2019). These methods use machine-learning models to automatically identify text or IFC data patterns that support information integration. However, existing machine learning-based information integration methods mostly rely on traditional semantic representations, such as the Word2vec (Zhou and El-Gohary 2020) and the global vectors for word representation (GloVe) (Zhang and El-Gohary 2019), and have not exploited the context-aware semantic representations generated using the pretrained transformer-based language models, which have recently achieved the state-of-the-art performance in various downstream natural language processing (NLP) tasks. There is, thus, a missing opportunity to leverage these pretrained language models to learn the semantic representations of the IFC and regulatory concepts for BIM-regulatory information integration.

To address this gap, this paper proposes an automated IFC-regulatory information integration approach based on the semantic representations generated by pretrained transformer-based language models that are finetuned using domain-specific corpora. The proposed approach consists of five main steps: data preparation, pretrained language model finetuning, semantic concept representation, semantic similarity-based concept alignment, and evaluation. The proposed approach was implemented and tested in integrating regulatory concepts from multiple building codes with IFC concepts from the IFC4 schema.

BACKGROUND

BIM-regulatory Information Integration. BIM-regulatory information integration aims to align natural-language regulatory information (e.g., regulatory concepts), which is typically extracted from requirements, with IFC information (e.g., IFC concepts such as IFC entities, properties, or enumerated property values). Existing efforts have focused on hardcoding using modeling languages (e.g., Yurchyshyna and Zarli 2009) or developing dictionaries (e.g., BuildingSMART 2020a), ontologies (e.g., Yurchyshyna and Zarli 2009), or rules (e.g., Pauwels et al. 2011) for mapping regulatory concepts to IFC concepts, either manually or semiautomatically. Despite the state-of-the-art performance achieved by these methods, they still require significant manual effort. Also, they often lack the flexibility to deal with the changes in the BIMs or the regulatory documents and are difficult to generalize to different types of regulatory documents (e.g., building code, energy conservation code, and accessibility standards) or different BIMs (e.g., BIMs of different levels of development) (Zhou and El-Gohary 2020).

A few research efforts have explored the use of machine-learning models such as classification models for BIM-regulatory information alignment for supporting ACC. For example, Zhang and El-Gohary (2016) developed a hybrid rule and ML-based method to extend the IFC schema. Zhou and El-Gohary (2018, 2020) further computed concept similarities based on the semantic representations (i.e., Word2vec) of concepts and explored different types of supervised learning algorithms and features in classifying the relationship between regulatory and IFC concepts.

Transformer-based Language Models. The transformer-based model architectures are deep learning models that consist of blocks – the transformers – that use attention mechanisms (Vaswani 2017) to model text data, and especially learn the contextual dependencies between words in the

text. Compared to other deep learning models that were predominately used for NLP tasks, transformer-based architectures have improved both the language modeling performance, especially in dealing with long term dependencies in the text, and the computational efficiency in model training, by (1) replacing the recurrent or convolutional neural networks with the attention-based transformers (Vaswani 2017); and (2) incorporating deeper model structures (e.g., millions of model parameters and several stacked model layers). The transformer-based model architectures enable pretraining language models on large open-domain corpora (e.g., Wikipedia). The pretrained transformer-based language models can then be finetuned on smaller, domain- or task-specific text data for downstream NLP tasks, such as sequence labeling, machine translation, and question answering (Vaswani 2017, Devlin et al. 2018, Radford et al. 2019). Two main categories of transformer-based language models have been proposed, pretrained, and implemented in solving these tasks: autoregressive models, such as the generative pre-trained transformer (GPT) series [e.g., GPT-2 (Radford et al. 2019)] by OpenAI, and autoencoding models, such as the bidirectional encoder representations from transformers (BERT) (Devlin et al. 2018) by Google and its variants.

PROPOSED APPROACH FOR BIM-REGULATORY INFORMATION INTEGRATION

The proposed approach for BIM-regulatory information integration consists of five main steps, as shown in Figure 1: regulatory and IFC data preparation, pretrained language model finetuning, semantic concept representation, semantic similarity-based concept alignment, and evaluation.

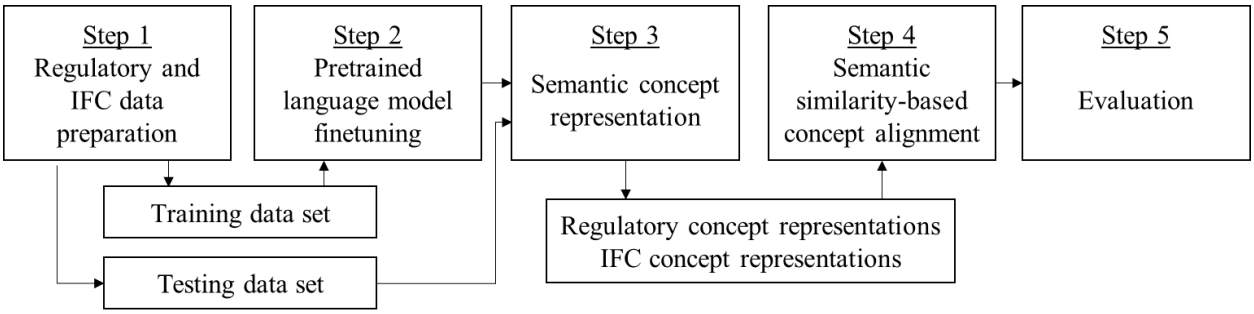

Figure 1. Proposed approach for BIM-regulatory information integration.

Regulatory and IFC Data Preparation. Two sets of data were prepared for training and testing. The training data set consists of 50,000 sentences and sentence fragments collected from various types of regulatory documents from the architecture, engineering, and construction (AEC) domain, including building codes, specifications, and standards. The sentences and sentence fragments were cleaned and tokenized, and were stored in a single TXT file. The training data set was later used for pretrained language model finetuning.

The testing data set consists of two parts: the regulatory and IFC data sets. The regulatory data set consists of 120 sentences collected from three types of regulatory documents: the International Building Code (IBC), International Energy Conservation Code (IECC), and Americans with Disabilities Act Standards for Accessible Design (ADA Standards). The sentences were cleaned and tokenized, and a total of 200 regulatory concepts were manually extracted from the sentences and mapped to IFC concepts contained in the IFC data set by three experts. The IFC concept data set, which consists of IFC concepts (e.g., IFC entities, properties, and enumerated property values) and the canonical forms of these concepts, were manually prepared based on the

buildingSMART International standards and supporting documentation on IFC4 (BuildingSMART 2020b). A canonical form is a natural-language description of an IFC concept. For example, the canonical form of the IFC entity IfcDoor is "door". Table 1 shows example IFC concepts, their canonical forms, and their related IFC concepts. The testing data set was later used for semantic concept representation, semantic similarity-based concept alignment, and evaluation.

Table 1: Example IFC Concepts in the Testing Data Set

IFC concept	IFC concept type	Canonical form	Example(s) of related IFC concepts
IfcDoor	Entity	door	• FireRating (via Pset_DoorCommon) • SINGLE_SWING_LEFT (via IfcDoorTypeOperationEnum) • OverallHeight (via IfcDoor attributes)
FireRating	Property	fire rating	• IfcDoor (via Pset_DoorCommon) • IfcStair (via Pset_StairCommon) • IfcWall (via Pset_WallCommon)
SINGLE_SWING_LEFT	Enumerated property value	single left swinging door	• IfcDoor (via IfcDoorTypeOperationEnum)

Pretrained Language Model Finetuning. Two types of transformer-based deep learning model architectures – BERT and GPT2 – were selected for learning the semantic representations of IFC and regulatory concepts. The pretrained base and uncased BERT and the pretrained base GPT2, both of which are accessible via the transformers library (Wolf et al. 2020), were finetuned on the training regulatory data set for capturing domain-specific text patterns. Three training practices were followed: (1) the sentences and sentence fragments were encoded using the BERT and GPT2 tokenizers, respectively; (2) the pretrained BERT was finetuned using the masked language modeling loss and the pretrained GPT2 was finetuned using the casual language modeling loss; and (3) the finetuning was stopped early to prevent potential overfitting.

Semantic Concept Representation. For each regulatory concept or canonical form of an IFC concept, the semantic concept representation was generated following three steps: word encoding, context-aware word representation generation, and concept representation construction. First, the sequence of words corresponding to the canonical form or the regulatory concept was encoded using the tokenizers of the pretrained language models. Second, the encoded sequence was fed into the models for computing the outputs of the hidden layers, which were used as the context-aware word representations of the sequence of words. Third, for each canonical form, the concept representation was constructed by averaging the word representations of the words contained in the canonical form or the regulatory concept. For each regulatory concept, two semantic representations were constructed – the complete and core semantic representations – by averaging the representations of all the words contained in the concept and directly using the representation of the last word in the concept, respectively. The size of each generated word or concept representation is 768, which equals the hidden-layer size of the base and uncased BERT and the base GPT2 (i.e., the number of neurons in the attention and output layers of the transformers in both models is 768).

Semantic Similarity-based Concept Alignment. Each regulatory concept was determined as aligned to an IFC concept (in the IFC concept data set) or not aligned (to any of the IFC concepts in the data set), following three main steps, as shown in Figure 2: core semantic similarity assessment, complete semantic similarity assessment, and threshold-based alignment. First, for each IFC concept, the semantic similarities between the core semantic representation and the semantic representations of all the canonical forms were calculated to obtain the IFC concept that has the maximum similarity. Second, the IFC concept that has the maximum similarity with the complete semantic representation was obtained similarly. Third, if the larger similarity of the two exceeds the information integration threshold, the regulatory concept is aligned to the IFC concept with this larger similarity; otherwise, the regulatory concept is unaligned.

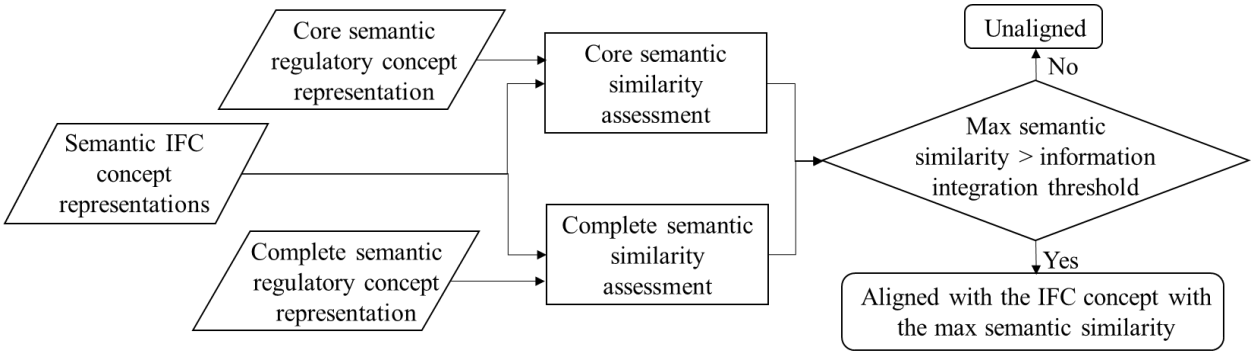

Figure 2. Semantic similarity-based concept alignment.

For each pair of a canonical form of an IFC concept and a regulatory concept, the semantic similarity was defined as the cosine similarity between the corresponding pair of semantic concept representations, as per Eq. (1), where $\mathbf{S_c}$ is the semantic representation of the canonical form of an IFC concept c and $\mathbf{S_r}$ is the semantic representation of the regulatory concept r.

$$Similarity\ (c,r) = \frac{\mathbf{S_c} \cdot \mathbf{S_r}}{\|\mathbf{S_c}\|\|\mathbf{S_r}\|} \quad (1)$$

Evaluation. The external evaluation metric, information integration accuracy (Zhang and El-Gohary 2016), was calculated, as per Eq. (2), where m is the number of regulatory concepts correctly aligned to their corresponding IFC concepts or not aligned to any IFC concepts according to the testing data set and N is the total number of regulatory concepts contained in the testing data set.

$$Accuracy = \frac{m}{N} \quad (2)$$

PRELIMINARY EXPERIMENTS AND RESULTS

Comparison of Different Pretrained Language Models. The two selected pretrained language models – the base and uncased BERT and the base GPT2 (with and without finetuning on the training regulatory text data) – were compared in terms of information integration accuracy. The experimental results, as shown in Table 2, indicate that (1) there was no significant difference

between the performances of the two types of pretrained language models; and (2) finetuning of the pretrained language models improved their performances.

Table 2: Pretrained Language Models and BIM-regulatory Information Integration

Metric	Pretrained language model			
	BERT (without finetuning)	GPT2 (without finetuning)	BERT (with finetuning)	GPT2 (with finetuning)
Information integration accuracy	0.74	0.75	0.80	0.80

Comparison of Different Semantic Representations. The semantic representations generated using the pretrained language models were compared with GloVe in terms of information integration accuracy. The experimental results, as shown in Table 3, indicate that the former performs better, because the pretrained language models captured the contextual semantic representations of the concepts, while GloVe only captured fixed, global semantic representations.

Table 3: Semantic Representations and BIM-regulatory Information Integration

Metric	Semantic representation	
	GloVe	Pretrained language models
Information integration accuracy	0.77	0.80

Comparison of Different Concept Alignment Thresholds. Three different information alignment thresholds were tested and compared in terms of information integration accuracy: 50%, 70%, and 90%. The experimental results, as shown in Table 4, indicate that the 70% threshold led to the best performance.

Table 4: Concept Alignment Threshold and BIM-regulatory Information Integration

Metric	Information alignment threshold		
	50%	70%	90%
Information integration accuracy	0.76	0.80	0.68

Error Analysis. Two types of errors were identified based on the experimental results. First, long and complex regulatory concepts (e.g., "side-hinged opaque door assembly") were not correctly aligned/unaligned, possibly because the averaging of word representations corresponding to these concepts reduced the semantic representations to be less discriminative in similarity-based concept alignment. Second, regulatory concepts that have words rarely occuring in the training data set or words very often occurring (e.g., stop words) were not correctly aligned/unaligned, possibly because the semantic representations were not able to capture the context-aware semantics that is discriminative in similarity-based concept alignment.

CONCLUSION

This paper proposed a new machine learning-based approach for semantic representation learning and information integration of building information modeling and regulatory concepts. First, the pretrained transformer-based language models (i.e., the base and uncased BERT and

the base GPT2) were finetuned on domain-specific regulatory text data. Second, for each regulatory concept and canonical form of an IFC concept, semantic concept representations were generated using the pretrained language models. Third, each regulatory concept was determined as aligned to an IFC concept or unaligned based on the semantic similarity between its semantic concept representations and these of the canonical forms of the IFC concepts. The proposed approach achieved an information integration accuracy of 80% on 200 regulatory concepts and IFC concepts from the IFC4 schema, indicating promising performance.

This paper contributes to the body of knowledge in two primary ways. First, the paper leverages pretrained transformer-based language models for capturing the semantic representations of IFC and regulatory concepts. Second, the initial experimental results show that the funetuning of pretrained language models and the information integration threshold can greatly impact the performance of the proposed approach.

In future work, first, the authors plan to improve the performance of the BIM-regulatory information integration by (1) incorporating the graph structures of the regulatory concepts in the requirements and IFC entities in the BIMs to allow complex concept alignment; (2) testing and comparing other state-of-the-art pretrained language models; and (3) including more IFC and regulatory concepts. Second, the authors plan to integrate the proposed BIM-regulatory information integration approach with machine learning-based regulatory information extraction and transformation approaches and compliance reasoning mechanisms in a machine learning-driven, fully automated compliance checking system.

ACKNOWLEDGEMENT

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.

REFERENCES

buildingSMART (2020a). buildingSMART Data Dictionary. <http://bsdd.buildingsmart.org/#peregrine/about> (Apr. 15, 2021).

buildingSMART (2020b). buildingSMART International Standards Server. <https://standards.buildingsmart.org> (Apr. 01, 2021).

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.

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), 506-518.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). "Language models are unsupervised multitask learners." *OpenAI blog*, 1(8), 9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017). "Attention is all you need." arXiv preprint arXiv:1706.03762.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S. and Louf, R. (2020). "Transformers: State-of-the-art natural

language processing." *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45.
Wu, J. and Zhang, J. (2019). "New automated BIM object classification method to support BIM interoperability." *J. Comput. Civ. Eng.*, 33(5), p.04019033.
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), 1084-1098.
Zhang, J., and El-Gohary, N. (2016). "Extending building information models semiautomatically using semantic natural language processing techniques." *J. Comput. Civ. Eng.*, 10.1061/(ASCE)CP.1943-5487.0000536, C4016004.
Zhang, R. and El-Gohary, N. (2019). "A Machine-Learning Approach for Semantic Matching of Building Codes and Building Information Models (BIMs) for Supporting Automated Code Checking." *In International Congress and Exhibition Sustainable Civil Infrastructures*, Springer, Cham, 64-73.
Zhou, P., and El-Gohary, N. (2018). "Automated matching of design information in BIM to regulatory information in energy codes." *In Construction Research Congress 2018: Construction Information Technology*, ASCE. doi: 10.1061/9780784481264.008
Zhou, P., and El-Gohary, N. (2020). "Semantic information alignment of BIMs to computer-interpretable regulations." *Advanced Engineering Informatics*, https://doi.org/10.1016/j.aei.2020.101239.