A Machine-Learning Approach for Semantic Matching of Building Codes and Building Information Models (BIMs) for Supporting Automated Code Checking

Ruichuan Zhang(✉) and Nora El-Gohary

Department of Civil and Environmental Engineering,
University of Illinois at Urbana-Champaign, Urbana, IL, USA
{rzhang65, gohary}@illinois.edu

Abstract. Various automated code compliance checking (ACC) systems have been developed and used to check the compliance of building information models (BIMs) with building codes, to reduce the time, cost, and errors of the code compliance checking process. All these systems require some form of code-BIM matching – matching of the concept representations in the codes to those in the BIMs – which is a difficult task. Traditionally, semantic matching was conducted in a highly-manual manner. To address this problem, more recently, a limited number of efforts have proposed fully automated semantic matching methods, which mostly rely on matching annotations and/or rules developed by domain experts. Despite their relatively good performance, these methods are by nature difficult to generalize or scale up (e.g., the matching rules need to be updated, modified, or extended when switching from one type of code to another). There is, thus, a need for semantic matching approaches that are more generalizable and scalable. To address this need, this paper proposes a new, machine learning-based approach to automatically match the building-code concepts and relations to their equivalent concepts and relations in the Industry Foundation Classes (IFC). The proposed approach consists of five primary tasks: (1) prepare and process the training and testing data; (2) automatically identify the domain word embeddings by learning from a large corpus of building-code text and generate the final semantic representations by combining the domain and general word embeddings; (3) match the building-code concepts to the IFC elements; (4) match the building-code relations to the IFC relations; and (5) evaluate the performance of the proposed approach using accuracy. The proposed approach was implemented and tested on a number of chapters from the 2009 International Building Code (IBC) and the Champaign 2015 IBC Amendments. The preliminary results show that the proposed approach achieved an accuracy of 77% for matching building-code concepts to IFC elements, and 78% for matching building-code relations to IFC relations, indicating promising semantic matching performance.

H. Rodrigues et al. (Eds.): GeoMEast 2019, SUCI, pp. 64–73, 2020.
https://doi.org/10.1007/978-3-030-34216-6_5

1 Introduction

To reduce the time, cost, and errors of compliance checking, various automated code compliance checking (ACC) systems have been developed and used to check the compliance of building information models (BIMs) with building codes. These systems have used different methods for information representation and code checking, and have achieved different levels of automation and performance. However, all of them require some form of code-BIM matching – matching of the concept representations in the codes to those in the BIMs. A certain level of matching can be conducted by simply matching natural language words and/or searching through the domain ontology (e.g., match "beam" to "IfcBeamTypeEnum – Beam"). However, it is difficult to match the regulatory information in building-code concepts represented by phrases and clauses, and building-code relations represented by verbs and/or adjectives, to Industry Foundation Classes (IFC) concepts (e.g., "return through" and "detoxification compound facilities and spaces" each cannot be directly matched to an IFC concept). Thus, to ensure the performance of the ACC systems, there is a need to develop an information matching approach that is capable to deal with regulatory information carried in natural language with diversified syntactic and semantic patterns.

In many cases, semantic matching was conducted in a highly-manual manner. With the increasing opportunities and needs for automation, more recent ACC efforts have, instead, proposed semi-automated (e.g., using machine learning algorithms to identify candidate matches, and requiring a human user/expert to verify these matches) (Zhang and El-Gohary 2016). Most recently, a limited number of efforts have also proposed fully automated semantic matching methods, which mostly rely on matching annotations and/or rules developed by domain experts (Zhou and El-Gohary 2018). Despite their relatively good performance, these methods are by nature difficult to generalize or scale up – when switching from one type of code to another, or from one chapter to another in the same code, the matching rules might need to be updated, modified, or extended. There is, thus, a need for semantic matching approaches that are more generalizable and scalable.

To address this need, this paper proposes a new, data-driven approach to automatically match the building-code concepts and relations to their equivalent concepts and relations in the IFC. The proposed approach consists of five primary tasks: (1) prepare and process the training and testing data; (2) automatically identify the domain word embeddings of the building-code concepts and relations by learning from a large corpus of building-code text and generate the final semantic representations by combining the domain and general word embeddings; (3) match the building-code concepts to the IFC elements using a similarity-based method; (4) match the building-code relations to the IFC relations using a supervised learning-based method; and (5) evaluate the performance of the proposed approach using accuracy.

2 Background

2.1 Semantic Matching

Semantic matching aims to identify the information that is semantically related (Fernández et al. 2011). Many research efforts have been undertaken to match information from sources such as text and information models other than building information models (BIMs) to the information from BIMs. For example, Cemesova et al. (2015) proposed PassivBIM to integrate the geometric and building fabric information from BIMs with the energy information in building performance simulation (BPS) models. Karan et al. (2015) used a semantic web-based method to identify the common entities among BIMs and geographic information (GIS) systems. Zhang and El-Gohary (2016) proposed a semi-automated learning-based method for matching the regulatory concepts and relations extracted from building codes to their most-related IFC concepts (e.g., equivalent concept, subconcept, superconcept) and relations for supporting ACC. Afsari et al. (2017) proposed ifcJSON representations to map information in the IFC data format to information in the JSON data format for facilitating web-based BIM data exchange. Zhou and El-Gohary (2018) proposed a rule-based method for matching the semantic information elements extracted from energy codes to the IFC concepts and relations for supporting energy code compliance checking.

The majority of semantic matching methods require that the semantic similarities between the concepts and relations are first established. To assess the semantic similarities between the concepts and relations – which are in the form of natural language, those concepts and relations need to be first represented in computer-processible semantic representations. Word embeddings is one of the most widely used semantic representations of natural language data. A word embedding is a vector representation of the word in a specific context (e.g., building code) (Mikolov et al. 2013). Word embeddings have been used for solving numerous text analytics tasks both in the computational linguistic domain [e.g., social media text classification (Xiao et al. 2018), semantic discovery (Yao et al. 2018)] and in the construction domain [e.g., building-code requirement analytics (Zhang and El-Gohary 2019)].

2.2 Industry Foundation Classes

The Industry Foundation Classes (IFC) data format aims to describe, represent, exchange, and share information typically used in the Architecture, Engineering, and Construction (AEC) domain, and is the most commonly used format of building information modeling (buildingSMART 2019). The IFC data format defines an object-based information model consisting of IFC elements and IFC relations. An IFC element is a physically existent component of a project in BIM (buildingSMART 2019). The most important IFC elements include the spatial structure elements (e.g., IfcSpace) and the building elements (e.g., IfcDoor). An IFC relation describes how the IFC elements are related to each other. For example, the "spatial composition" relation describes the case where a spatial structure element decomposes into other IFC elements. However, the IFC concepts do not correspond to the concepts and relations used in the building codes, which creates a major barrier for ACC. It makes the process of matching the building-code concepts and relations to the IFC elements and relations complex and challenging.

3 Proposed Machine-Learning Approach for Semantic Matching of Building Codes and Building Information Models

The proposed approach consists of five main steps, as shown in Fig. 1: (1) data preparation and preprocessing, (2) semantic representation generation, (3) similarity-based matching of building-code concepts to IFC elements, (4) supervised learning-based matching of building-code relations to IFC relations, and (5) method evaluation.

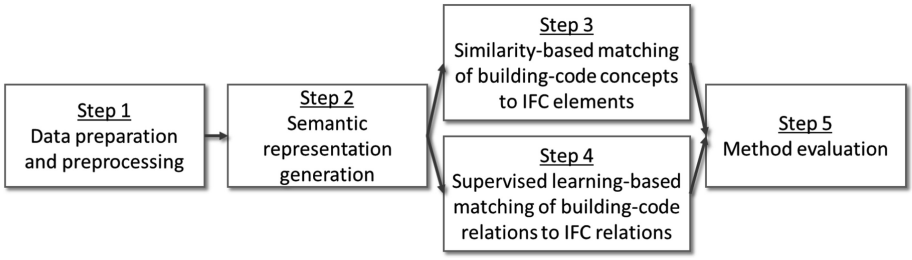

Fig. 1. Proposed machine-learning approach for semantic matching of building-code concepts and relations to Industry Foundation Classes (IFC) elements and relations

Step 1: Data Preparation and Preprocessing
For evaluating the matching of building-code concepts and relations, 80 sentences were selected from the 2009 IBC and the Champaign 2015 IBC Amendments. The concepts and relations in these sentences were manually extracted, resulting in a total of 97 building-code concepts and 73 building-code relations (including the two building-code concepts associated by this building-code relation, Concept A and Concept B). Each word in the names of the concepts and the relations was lowercased and singularized. All the concepts and relations were manually matched to the equivalent or super IFC elements or IFC relations, resulting in matching the 97 concepts to a total of 24 IFC elements and the 73 relations to a total of six IFC relations, as shown in Tables 1 and 2, respectively. The six relations include five original IFC relations, in addition to an added sixth relation, "complex relation", which was added to represent the case where a single building-code relation needs to be matched to multiple IFC relations. Each IFC element was further transformed into a canonical form – which is a lowercased English word, phrase, or sentence – for the purpose of semantic similarity assessment (Step 3). The transformation was conducted following three steps: (1) removing the prefixes in the IFC elements (e.g., "Ifc"), (2) referring to the explanations provided by the IFC documentation, and (3) using engineering judgment. For example, "IfcDoor" was transformed into "door", and "IfcSpace" was transformed into "room, space, or unit".

Table 1. Industry Foundation Classes (IFC) elements used in the proposed semantic matching approach

Type of IFC elements	IFC elements
Spatial structure elements	IfcSite, IfcBuilding, IfcBuildingStorey, IfcSpace
Building elements	IfcBeam, IfcChimney, IfcColumn, IfcCovering, IfcCurtainWall, IfcDoor, IfcFooting, IfcMember, IfcPile, IfcPlate, IfcRailing, IfcRamp, IfcRoof, IfcSlab, IfcStair, IfcWall, IfcWindow
Furnishing elements	IfcFurniture
Transportation elements	IfcTransportElementTypeEnum – Escalator, IfcTransportElementTypeEnum – Elevator

Table 2. Industry Foundation Classes (IFC) relations used in the proposed semantic matching approach

IFC relations	Definitions
Spatial composition	A spatial structure element decomposes into other IFC elements
Spatial container	A spatial structure element contains other IFC elements
Product placement	An IFC element's location relative to another IFC element
Material constituent	An IFC element consists of a material element
Property	An IFC element has a property

Two types of data were prepared for generating the semantic representations (Step 2). For identifying the domain word embeddings, a corpus of 6,000 sentences from the 2009 IBC and the Champaign 2015 IBC Amendments were used to train an unsupervised learning algorithm. For the general word embeddings, the "pre-trained word embeddings" (Pennington et al. 2014) were used. Those word embeddings were learned from a large, cross-domain corpus, using the Glove algorithm, and thus can provide additional semantic information (Pennington et al. 2014) to enhance the robustness of the semantic representations.

Step 2: Semantic Representation Generation

The semantic representations of the building-code concepts and relations, and the IFC elements and relations, were generated based on word embeddings. The semantic representation generation step consists of two substeps: (1) training the learning algorithm for identifying the domain word embeddings, and (2) combining the domain and the general word embeddings. First, the unsupervised learning algorithm, word2vec, was trained on the domain-specific corpus of building-code sentences using the Gensim (Rehurek and Sojka 2010) built in Python, in order to identify the domain word embeddings. Second, for each word, the final word semantic representation was computed as the weighted average of the domain and general embeddings, in order to reflect both the domain-specific semantic meanings and the general semantic meanings of the word. The domain semantic weight ranges from 0 to 1, where 0 represents only

using general semantic meanings and 1 represents only using domain semantic meanings.

Step 3: Similarity-Based Matching of Building-Code Concepts to IFC Elements
The building-code concepts were matched to the IFC elements using a similarity-based method, which consists of two substeps: semantic similarity assessment and concept-element matching. First, the semantic similarities between the building-code concepts and the canonical forms corresponding to the IFC elements were computed. Two semantic similarities were proposed: phrase similarity and last-word similarity. Phrase similarity is defined as the cosine similarity between the phrase semantic representations of the building-code concept and the canonical form of the IFC element. A phrase semantic representation is formed by averaging the word semantic representations of all words in the building-code concept or the canonical form. Last-word similarity is defined as the cosine similarity between the semantic representation of the last noun in the building-code concept – which typically carries important information about building elements – and the phrase semantic representation of the canonical form of the IFC element. Second, the building-code concepts were matched to the IFC elements based on matching scores. For each pair of building-code concept and IFC element, the higher one of the phrase and last-word similarities was used as the matching score. For each building-code concept, the candidate IFC element having the highest matching score was selected as the match.

Step 4: Supervised Learning-Based Matching of Building-Code Relations to IFC Relations
The building-code relations were matched to the IFC relations using a supervised learning-based method, which consists of two substeps: semantic feature development and relation classification. First, four semantic features were selected: the phrase semantic representations of the relation, Concept A, and Concept B, and the lettercase of the words in Concept B. Similar to Step 3, the phrase semantic representations were computed as the average of the semantic representations of the words in the relations and the concepts. The fourth feature (i.e., the lettercase of the words in the object) is binary, indicating whether there is a capitalized word in the object. Second, a relation classification model was trained using the training data. Two types of classifiers were tested and compared: a multilayer perceptron (MLP) and a multiclass support vector machine (SVM) with a linear kernel. The trained relation classification model is able to take new features and predict the corresponding IFC relations automatically.

Step 5: Evaluation
The performances of matching building-code concepts to IFC elements and matching building-code relations to IFC relations were evaluated separately, both using accuracy (Olson and Delen 2008). Accuracy is defined as the proportion of the testing building-code concepts or relations that are correctly matched to their corresponding IFC elements or relations, in the entire testing building-code concepts or relations dataset.

4 Preliminary Experimental Results

4.1 Performances in Code-BIM Matching

The performance of the proposed approach is summarized in Table 5. Based on the testing results, the accuracy of matching the building-code concepts to the IFC elements is 77% and the accuracy of matching the building-code relations to the IFC relations is 78%. Examples of the correctly matched pairs of building-code concepts and IFC elements, and pairs of building-code relations (with associated concepts) and IFC relations, are shown in Tables 3 and 4, respectively.

Table 3. Example matched building-code concepts and Industry Foundation Classes (IFC) elements

Building-code concepts (A)	Matched IFC element (B)	Type of match (relation of B to A)
Horizontal sliding power-operated door	IfcDoor	Superconcept
Building	IfcBuilding	Equivalent concept
Permanently installed furnishing	IfcFurniture	Superconcept
Mezzanine	IfcFloor	Superconcept
Type A dwelling unit	IfcSpace	Superconcept

Table 4. Example matched building-code relations and Industry Foundation Classes (IFC) relations

Building-code relations (with associated concepts)	Matched IFC relations	Type of match
Have, dwelling unit, room	Spatial composition	Equivalent relation
With, room, furred ceiling	Spatial container	Equivalent relation
To, egress, exit	Product placement	Equivalent relation
Have, corridor, ceiling height	Property	Equivalent relation
Accessory to, area, area	Complex relation	Equivalent relation

For matching the building-code concepts to the IFC elements and matching the building-code relations to the IFC relations, different domain semantic weights were tested and compared, including 0 (using general word embeddings only), 0.25, 0.50, 0.75, and 1 (using domain word embeddings only), as shown in Table 5. The optimal performance for matching the building-code concepts to the IFC elements was achieved when the domain semantic weight was 0.25; and the optimal performance for matching the building-code relations to the IFC relations was achieved when the domain semantic weight was 0.50. Compared to using only either domain word embeddings or general word embeddings, the use of weighted word embeddings

(i.e., the proposed semantic representation) increased the accuracies by up to 22%, which indicates the benefit of integrating both domain-specific and cross-domain semantic information.

For matching the building-code relations to the IFC relations, the two tested classification algorithms (i.e., MLP and SVM) achieved different performances for different semantic weights, but achieved the same optimal performance when the domain semantic weight is 0.50, as shown in Table 5.

4.2 Error Analysis

Two main types of errors were identified based on the experimental results. First, for matching building-code concepts to IFC elements, the proposed method had errors when dealing with building-code concepts that are less frequently appearing in the building code, such as "casework", which appears only once in the entire IBC 2009. The generated domain word embeddings may not be able to capture the domain semantic meanings of those concepts. In future work, a larger, more diversified corpus of text from the construction domain could be used for training. Second, for matching the building-code relations to the IFC relations, the proposed method misclassified "spatial composition" as "spatial container" or "complex relation". In future work, more training data and features based on domain ontology could be used, in order to enhance the ability of the relation classification model to distinguish such relation types that are close or related.

Table 5. Performance of the proposed approach with different domain semantic weights

Domain semantic weights	Accuracy of matching building-code concepts and Industry Foundation Classes (IFC) elements[a]	Accuracy of matching building-code relations and IFC Relations[a]	
		Multilayer perceptron	Multiclass support vector machine
0	76%	61%	56%
0.25	**77%**	69%	74%
0.50	71%	**78%**	**78%**
0.75	70%	69%	69%
1	66%	65%	69%

[a]Bolded font indicates the highest performance

5 Conclusions

This paper proposed a new machine learning-based approach for matching semantic information in building codes and building information models for supporting automated compliance checking (ACC), by separately matching the building-code concepts to the IFC elements and matching the building-code relations to the IFC relations. First, the semantic representations were generated by combining the domain word embeddings and the general word embeddings to reflect both domain-specific and cross-domain semantic information, in order to improve both accuracy and scalability of the proposed approach. Second, a similarity-based method was proposed to match the building-code concepts to the IFC elements. Third, a supervised learning-based method was proposed to match the building-code relations to the IFC relations. The proposed approach achieved a 76% accuracy of matching the building-code concepts to the IFC elements, and a 78% accuracy of matching the building-code relations to the IFC relations.

This paper contributes to the body of knowledge in two primary ways. First, the paper proposed a new way to model the semantic meanings of the domain-specific text by first generating the domain word embeddings and then combining both the domain and the general word embeddings. The proposed approach makes use of both domain and general semantic representations in semantic matching, and thus has potentially better scalability in dealing with different types of building codes. Second, the initial experimental results show that the proposed semantic representation successfully captured the semantic meanings of both building-code concepts and relations, and IFC elements and relations, in both similarity-based and supervised learning-based semantic matching tasks.

In their future work, the authors first plan to improve the information matching by including more IFC elements (e.g., IfcSanitaryTerminal), the properties of the IFC elements (e.g., Pset_DoorCommon – IsExternal), the subconcepts of the IFC elements [e.g., revolving door (a subconcept of IfcDoor)], and more IFC relations (e.g., element filling); and including building-code concepts and relations described in complex phrases (e.g., occupant evacuation elevator lobby) and sentences. Second, the authors will explore further ways to improve the performance of the proposed information matching approach, including using more training data for the domain word-embedding generation, annotating more training data for relation classification, and exploring different data similarities for matching the building-code concepts to the IFC elements, and different supervised learning algorithms for matching the building-code relations to the IFC relations. Third, and most importantly, the authors plan to integrate the proposed information matching approach with machine learning-based information extraction and transformation approaches, with an aim to develop a fully automated, and highly scalable ACC system.

Acknowledgments. 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

Zhang, J., El-Gohary, N.: Extending building information models semiautomatically using semantic natural language processing techniques. J. Comput. Civ. Eng. ASCE (2016). https://doi.org/10.1061/(asce)cp.1943-5487.0000536

Zhou, P., El-Gohary, N.: Automated matching of design information in BIM to regulatory information in energy codes. In: Construction Research Congress 2018: Construction Information Technology, ASCE (2018)

Cemesova, A., Hopfe, C.J., Mcleod, R.S.: PassivBIM: enhancing interoperability between BIM and low energy design software. Autom. Construct. (2015). https://doi.org/10.1016/j.autcon.2015.04.014

Karan, E.P., Irizarry, J., Haymaker, J.: BIM and GIS integration and interoperability based on semantic web technology. J. Comput. Civ. Eng. ASCE (2015). https://doi.org/10.1061/(asce)cp.1943-5487.0000519

Afsari, K., Eastman, C.M., Castro-Lacouture, D.: JavaScript Object Notation (JSON) data serialization for IFC schema in web-based BIM data exchange. Autom. Construct. (2017). https://doi.org/10.1016/j.autcon.2017.01.011

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E.: Semantically enhanced information retrieval: an ontology-based approach. J. Web Semantics (2011). https://doi.org/10.1016/j.websem.2010.11.003. Special Issue on Semantic Search

Yang, X., Macdonald, C., Ounis, I.: Using word embeddings in twitter election classification. Inf. Retrieval J. (2018). https://doi.org/10.1007/s10791-017-9319-5

Yao, Z., Sun, Y., Ding, W., Rao, N., Xiong, H.: Dynamic word embeddings for evolving semantic discovery. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM (2018). https://doi.org/10.1145/3159652.3159703

Zhang, R., El-Gohary, N.: A machine learning-based approach for building code requirement hierarchy extraction. In: Proceedings of the 7th CSCE International Construction Specialty Conference (Jointly with Construction Research Congress), CSCE (2019)

Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, ACL (2014). https://doi.org/10.3115/v1/d14-1162

Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA (2010)

Olson, D.L., Delen, D.: Advanced Data Mining Techniques. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-76917-0

buildingSMART: Industry Foundation Classes, Version 4 - Addendum 2, 15 June 2019. http://www.buildingsmart-tech.org/ifc/IFC4/Add2/html/