Extending Building Information Models Semi-Automatically Using Semantic Natural Language Processing Techniques

Jiansong Zhang[1] and Nora M. El-Gohary[2]

[1]Graduate Student, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61801; PH (217) 607-6006; FAX (217) 265-8039; email: jzhang70@illinois.edu
[2]Assistant Professor, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61801; PH (217) 333-6620; FAX (217) 265-8039; email: gohary@illinois.edu

ABSTRACT

Automated compliance checking (ACC) of building designs requires automated extraction of information from building information models (BIMs). However, current Industry Foundation Classes (IFC)-based BIM models provide limited support for ACC, because they lack the necessary information that is needed to perform compliance checking (CC). In this paper, we are proposing a new approach for extending the IFC schema to incorporate CC-related information, in a semi-automated and objective manner. Our method utilizes semantic natural language processing (NLP) techniques to extract concepts and relations from documents that are related to CC (e.g. building codes). We utilize pattern-matching-based rules for the extraction. We use three types of features in the matching patterns: part-of-speech tags, dependency relations, and term sequence numbers in a sentence. The selected concepts and relations are then automatically encoded into the EXPRESS-language-represented IFC schema. The automated encoding in EXPRESS is enabled using a set of mapping rules. To evaluate our proposed approach, we compared the concepts and relations that we automatically extracted from the International Building Code 2006 to extend the IFC schema with a manually-developed gold-standard, and evaluated the results in terms of precision and recall. We achieved higher than 90% precision and recall, which shows that our approach is promising.

INTRODUCTION

The use of building information models (BIMs) is supposed to support automated compliance checking (ACC) of building designs (Liebich 2009). However, current Industry Foundation Classes (IFC)-based BIMs provide limited

support for ACC, because they lack the necessary information that is needed to perform compliance checking (CC). To address this barrier, a number of existing research and software development efforts (e.g. Tan *et al*. 2010; Niemeijer *et al*. 2009) proposed different ways for extending the IFC schema to cover more information for supporting CC. These extension methods, however, are mostly ad-hoc and subjective (relying on subjective extensions by individual researchers); and the resulting extended models are usually still missing essential CC-related information (Niemeijer *et al*. 2009). To address this theoretical and practical gap, in this paper, we propose a new approach for extending the IFC schema with regulatory requirement information, objectively and semi-automatically. Our method utilizes semantic natural language processing (NLP) techniques and pattern-matching-based rules to automatically extract concepts and relations from construction regulatory documents (textual documents) to extend the IFC schema.

BACKGROUND

Building information models and Industry Foundation Classes. A building information model (BIM) is a "digital representation of physical and functional characteristics of a facility" (NBIMSPC 2013). A BIM could support various functions such as clash detection (Leite *et al*. 2011), 4D visualization, and ACC (Liebich 2009). The Industry Foundation Classes (IFC) has been widely used as the specification for BIM and has been registered with ISO as an official international standard (BuildingSMART 2013). Due to the broad coverage of domains (architectural, structural, construction, facility management, etc.) and project phases in the IFC schema, the IFC schema lacks rigorous, formally, and semantically-defined concepts and relations for some specific sub-domains or processes (Venugopal *et al.* 2012). In this regard, researchers have proposed various ways to extend the IFC schema with concepts and relations for specific purposes such as ACC. For example, Tan *et al*. (2010) defined the Extended Building Information Model (EBIM) to incorporate building hygrothermal performance simulation results (from a simulation software) into an XML-language-represented IFC schema; Niemeijer *et al*. (2009) proposed to use abstract syntax trees of constraints to extend the IFC schema with missing concepts and relations; and the Singapore CORENET project extended the IFC schema using FORNAX (i.e. a C++ library to derive new data and generate extended views of IFC data) objects (Eastman *et al*. 2009). However, to avoid inconsistency and incompleteness in the extension of the IFCs, a more objective and automatic way of extending the IFC schema is needed.

Natural language processing and dependency relations. Natural language processing (NLP) is a field in artificial intelligence and linguistics that aims at enabling computers to understand and process natural languages (i.e. text and speech)

in a human-like manner. Example applications of NLP include document classification, information extraction, and machine translation (Marquez 2000; Salama and El-Gohary 2013). Sentence structural analysis aims at enabling such NLP applications. There are mainly two ways of conducting sentence structural analysis: using constituency grammar or using dependency grammar (Covington 2001). Constituency grammar originates from phrase structure grammar introduced by Chomsky (1956). Constituency grammar is based on constituency relations, which captures the relations between terms in a sentence in a hierarchical breaking-down manner (Figure 1). Constituency grammar is the basis for formal language theory (Covington 2001). Dependency grammar is based on dependency relations. Dependency relations capture the relations between terms in a sentence using pairwise linkages (Figure 1). For information extraction, the use of dependency relations have shown to require less extraction rules in comparison to the use of constituency relations (Zhang and El-Gohary 2012).

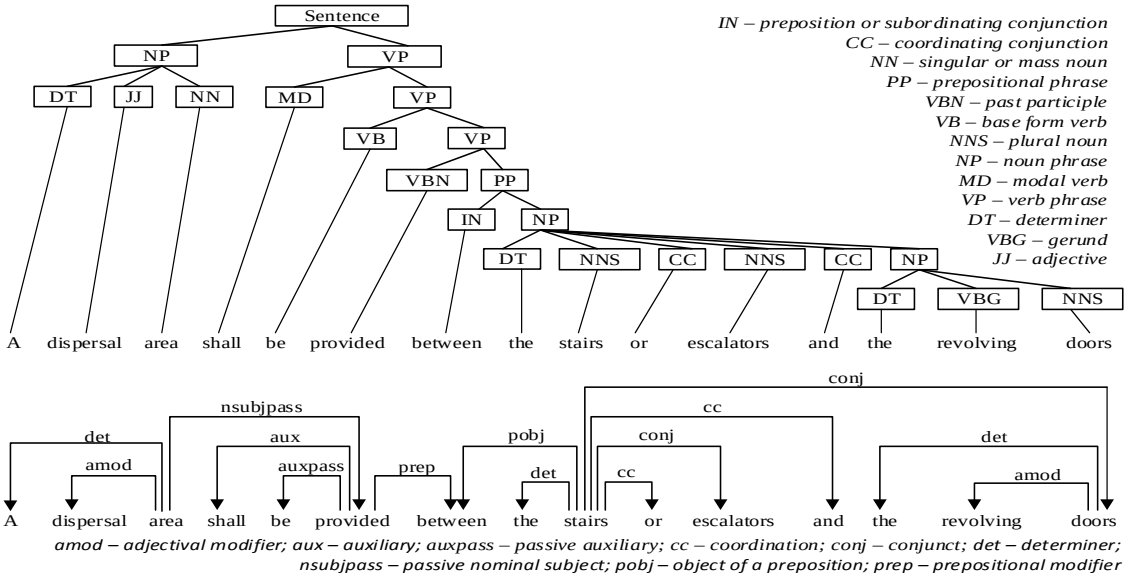

Figure 1. Constituency relations (top) and dependency relations (bottom) of a sentence.

PROPOSED SEMI-AUTOMATED IFC SCHEMA EXTENDING METHOD

We are proposing a three-phase method for semi-automatically extending the IFC schema with concepts and relations from regulatory documents (Figure 2).

Phase 1 – Semantic rule-based extraction. This phase aims at extracting all potential concepts and relations from regulatory documents that are related to the concepts of the IFC schema. This phase has two main inputs: 1) concepts from the

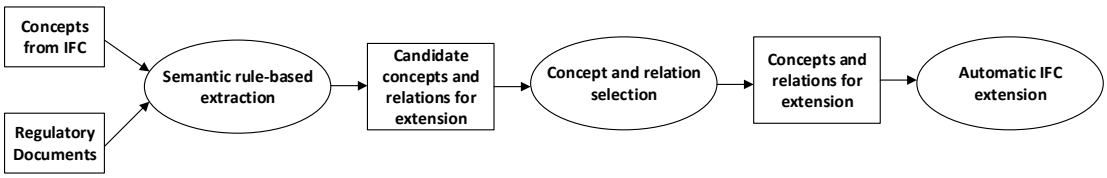

Figure 2. Proposed IFC schema extending method.

IFC schema, and 2) regulatory documents. The concepts from the IFC schema (we call them "seeds" hereafter) are processed one by one for the extraction of related concepts and relations. We utilize pattern-matching-based rules for extraction. In each rule, the left-hand-side defines the pattern to be matched and the right-hand-side defines which part of the matched pattern to extract. We utilize three types of features in the matching patterns: 1) part-of-speech (POS) tags, 2) dependency relations, and 3) term sequence numbers in a sentence. POS tags are the labels assigned to each word of a sentence indicating their lexical or functional categories. Example POS tags include DT (determiner), VBN (past participle), and CC (coordinating conjunction), etc. (Bellegarda 2010). Dependency relations are the pairwise "linkages" between terms in a sentence indicating their grammatical relations. Example dependency relations include amod (adjective modifier), prep (prepositional modifier), and pobj (object of a preposition), etc. (Marneffe and Manning 2013). Term sequence numbers in a sentence indicate the count of terms in the sentence up to the current term. A semantic model (ontology) about the types of relations between target concepts (to extract) and seeds guides the development of our extraction rules. Figure 3 shows the small ontology that we defined and used for extracting concepts and relations from construction regulatory documents that are related to "seeds". In the ontology, 'relation' has three sub-concepts: 1) 'subtype relation', 2) 'has-part relation', and 3) 'cross-concept relation'. A 'subtype relation' indicates that a concept is a sub-class of another concept. For example, "revolving door" is a subtype of "door". A 'has-part' relation indicates that the preceding concept contains the following concept as a part. For example, "window" has "sash" as a part. A 'cross-concept relation' defines any relation between concepts other than a 'subtype relation' or a 'has-part relation'. We defined two 'cross-concept relations': 'has-property relation' and 'constraint relation'. A 'has-property' relation indicates that the preceding concept has a property represented by the following concept. For example, "window" has "total area" as a property. A 'constraint relation' indicates that a concept is restricted by other concepts. There are two types of 'constraint relations': 'descriptive constraint relation' which defines a constraint relation through descriptive statements; and 'operative constraint relation' which defines a constraint relation through enforcing actions. There are two types of 'descriptive constraint relations': 'physical containing constraint relation' which defines the relation that a concept physically contains another concept (e.g. 'window in walls' defines the physical relation that the

"window" in description is contained in "wall"); and 'bounding range constraint relation' which represents the relation that the preceding concept is bounded by the following concepts (e.g. "windows between fabrication areas and corridors" constrain the range of "window" using the concepts "fabrication areas" and "corridors"). There are two types of 'operative constraint relations': 'action constraint relation' which represents the relation that the preceding concept exerts an action on the following concept (e.g. "windows comply with Section 715.5" applies the 'action constraint' "comply with Section 715.5" to "window"); and 'reversed action constraint relation' which represents the relation that the following concept exerts an action on the preceding concept (e.g. "windows constructed of approved materials" applies the 'reversed action constraint' "constructed of approved materials" on "window").

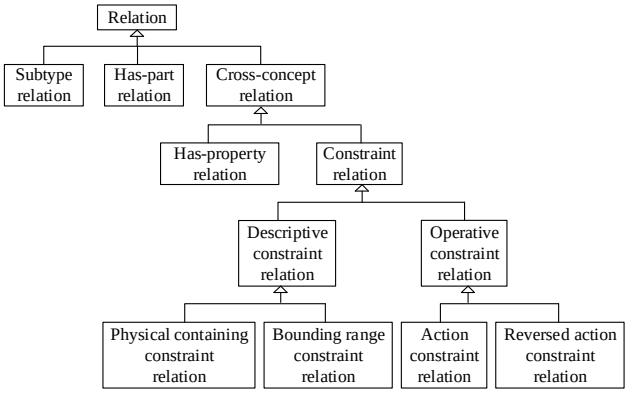

Figure 3. Our relation ontology.

Phase 2 – Concept and relation selection. This phase aims at filtering out the concepts and relations that we use to extend the IFC schema, from the concepts and relations extracted from Phase 1. We propose two methods for this task: 1) expert-judgment-based manual selection, which utilizes experts' knowledge and judgment to manually pick out the concepts and relations to use; and 2) semantic-algorithms-based automated selection, which utilizes semantic rule-based algorithms based on a detailed ontology to automatically filter out the concepts and relations that are compatible with the detailed ontology.

Phase 3 – Automatic IFC schema extension. This phase aims at automatically extending the IFC schema with the concepts and relations selected in Phase 2. We chose to use the EXPRESS-language-represented IFC schema for this extension. We utilized mapping-rule-based algorithms for this automated task. The rules map the selected concepts and relations (from Phase 2) into EXPRESS code that could be directly appended to the IFC schema. Figure 4 shows an example mapping rule.

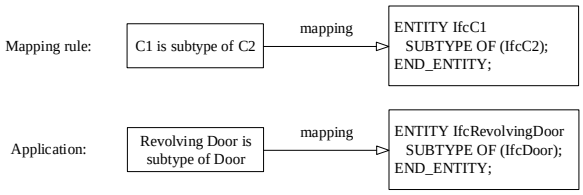

Figure 4. An example mapping rule and its application.

PRELIMINARY EXPERIMENTAL RESULTS AND ANALYSIS

The proposed method was tested on extending the IFC schema based on the International Building Code (IBC) 2006 (ICC 2006). We used the most updated version of the IFC schema – IFC4. Among the concepts in IFC4, we selected "window" as the seed for developing extraction rules and "door" as the seed for testing the developed extraction rules. We used python programming language (Python 3.2) to encode our extraction rules for testing. We developed subordinate functions and methods on-the-go in the python script as needed. We used POS tags (generated by the PCFG parser), typed dependency relations (Marneffe and Manning 2013), and term sequence numbers as features (all generated by the Stanford parser Version 3.3.0) in the matching patterns. The experimental results are shown in Table 1.

Extraction rule development. To develop the extraction rules, we first extracted all sentences in IBC 2006 that contain the seed "window". We selected the first 100 of such sentences, and developed our pattern-matching-based extraction rules based on them. We adopted an iterative testing-driven approach for the rule development: run all the extraction rules on the 100 sentences each time a new rule is added. If the newly-added rule affects any previous sentence(s) and adjusting the newly-added rule by itself could not avoid such effect, then all the rules need to be adjusted together. We utilized the relation ontology (Figure 3) to guide our rule development. Twenty pattern-matching-based extraction rules were developed based on the 100 sentences.

Extraction rule testing. To test the developed extraction rules, we first extracted all sentences in IBC 2006 that contain the seed "door". We randomly selected 100 sentences for testing. We developed a gold standard based on the 100 sentences. The gold standard contains 168 concept-relation combinations. We used concept-relation combination as the basic testing unit because each extracted concept connects to the seed through a relation without which the concept would not be meaningful. For example, "revolving door is_subtype_of door" is a concept-relation combination.

Evaluation. We evaluated the testing results in terms of precision, recall, and F1 measure. Precision is the number of correctly extracted concept-relation combinations divided by the total number of concept-relation combinations extracted. Recall is the number of correctly extracted concept-relation combinations divided by the total

number of concept-relation combinations that should be extracted (i.e. that are in the gold standard). F1 measure is the harmonic mean of precision and recall.

Table 1. Preliminary Experimental Results.

Number of concept-relation combinations in gold standard	168
Total number of concept-relation combinations extracted	174
Number of concept-relation combinations correctly extracted	158
Precision	0.91
Recall	0.94
F-Measure	0.92

The preliminary experimental results show more than 90% performance in all measures of precision, recall, and F1 measure. This indicates that our proposed concept-relation extraction method is promising. Through error analysis we recognized two main sources of error: 1) unseen patterns in extraction rule development, and 2) errors propagated from POS tagging. For future improvement, we plan to increase the number of sentences to use when developing our extraction rules, so that the probability of missing certain patterns could be reduced.

CONCLUSION AND FUTURE WORK

In this paper, the authors proposed a three-phase method for extending the IFC schema semi-automatically for supporting ACC. Our method utilizes semantic NLP techniques and pattern-matching-based rules to automatically extract candidate concepts and relations from construction regulatory documents to extend the IFC schema. We used three types of features in the pattern-matching-based rules: 1) POS tags, 2) dependency relations, and 3) term sequence numbers in a sentence. Evaluation in terms of precision, recall, and F1 measure against a gold standard shows a performance greater than 90% for all three measures. This shows that our proposed method is promising. In the future, we plan to increase the number of sentences used in rule development to further improve the performance. Also, we plan to test our method on more “seed” concepts.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Foundation (NSF). This material is based upon work supported by NSF under Grant No. 1201170. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of NSF.

REFERENCES

Bellegarda, J.R. (2010) "Part-of-Speech tagging by latent analogy." *IEEE J. Selected Topics in Signal Processing*, 4(6), 985-993.

BuildingSMART. (2013). "IFC overview summary."< http://www.buildingsmart-tech.org/specifications/ifc-overview> (Dec.01, 2013).

Chomsky, N. (1956). "Three models for the description of language." *IRE Transactions on Information Theory*, 113-124.

Covington, M.A. (2001). "A fundamental algorithm for dependency parsing." *Proc., 39th Annual ACM Southeast Conf.*, ACM, Athens, Georgia, 95-102.

Eastman, C., Lee, J., Jeong, Y., and Lee, J. (2009) "Automatic rule-based checking of building designs." *Autom. Constr.*, 18(8), 1011-1033.

International Code Council (ICC). (2006). "2006 International Building Code." <http://publicecodes.cyberregs.com/icod/ibc/2006f2/> (Dec. 02, 2013).

Leite, F., Akcamete, A, Akinci, B., Atasoy, G., and Kiziltas, S. (2011). "Analysis of modeling effort and impact of different levels of detail in building information models." *Autom. Constr*., 20, 601-609.

Liebich, T. (2009). "IFC 2x Edition 3 model implementation guide." buildingSMART International, 108-110.

Marneffe, M., and Manning, C. (2013). "Stanford typed dependencies manual." < http://nlp.stanford.edu/downloads/dependencies_manual.pdf> (Dec. 02, 2013).

Marquez, L. (2000). "Machine learning and natural language processing." *Proc., "Aprendizaje automatico aplicado al procesamiento del lenguaje natural"*.

National Building Information Model Standard Project Committee (NBIMSPC). (2013). "National BIM Standard – United States Version 2." <http://www.nationalbimstandard.org/faq.php#faq1> (Dec. 01, 2013).

Niemeijer, R.A., Vries, B.D., and Beetz, J. (2009). "Check-mate: automatic constraint checking of IFC models." In A Dikbas, E Ergen & H Giritli (Eds.), *Manag. IT in Constr. Manag. Constr. for Tomorrow*, CRC Press, London, UK, 479-486.

Salama, D.M., and El-Gohary, N.M. (2013). "Semantic text classification for supporting automated 801 compliance checking in construction." *J. Comput. Civ. Eng.*, accepted.

Tan, X., Hammad, A., and Fazio, P. (2010). "Automated code compliance checking for building envelope design." *J. Comp. in Civ. Eng*., 24(2), 203-211.

Venugopal, M., Eastman, C.M., Sacks, R., and Tezer, J. (2012). "Semantics of model views for information exchanges using the industry foundation classes schema." *Advanced Eng. Informatics*, 26, 411-428.

Zhang, J., and El-Gohary, N.M. (2012). "Extraction of construction regulatory requirements from textual documents using natural language processing techniques." *Proc., 2012 ASCE Intl. Conf. on Comput. Civ. Eng.*, ASCE, Reston, VA, 453-460.