5th International/11th Construction Specialty Conference
5e International/11e Conférence spécialisée sur la construction

Vancouver, British Columbia
June 8 to June 10, 2015 / *8 juin au 10 juin 2015*

A SEMANTIC SIMILARITY-BASED METHOD FOR SEMI-AUTOMATED IFC EXTENSION

Jiansong Zhang[1,2], and Nora El-Gohary[1]
[1] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, USA
[2] jzhang70@illinois.edu

Abstract: The Industry Foundation Classes (IFC) schema was designed as a comprehensive data schema to cover information of all phases of a building project and all disciplines of the AEC industry. But due to its limited coverage of details in certain subdomains, the IFC schema needs to be extended for many information processing tasks such as information extraction for automated regulatory compliance checking. Previous IFC extension efforts typically extended IFC in an ad-hoc and subjective manner. A more objective, standardized, and application-independent method for extending IFC is, thus, needed. To address this gap, a new method for extending the IFC schema objectively and semi-automatically is proposed. The proposed method utilizes a semantic relation-based concept matching algorithm to find concepts – from domain documents – to incorporate into the current IFC schema class hierarchy. It utilizes the hypernymy, hyponymy, and synonymy semantic relations. This paper focuses on presenting the proposed semantic relation-based concept matching algorithm: the ZESeM (Zhang and El-Gohary Semantic Matching) algorithm. The ZESeM algorithm was tested on processing concepts from Chapter 12 of the International Building Code 2006. Different semantic similarity computation methods were tested in combination with the proposed ZESeM algorithm. The ZESeM algorithm was evaluated based on adoption rate, which is the number of concepts found by the ZESeM algorithm that are adopted divided by the total number of concepts found by the ZESeM algorithm. An adoption rate of 85.8% was achieved. The proposed semantic relation-based concept matching algorithm offers a more efficient concept matching method for semi-automatically extending the IFC schema.

1 INTRODUCTION

Building projects must comply with various regulations, such as the International Building Code (IBC), the Americans with Disabilities Act, the federal Fair Housing Act, and the Occupational Safety and Health Administration regulations. Due to the large amount of requirements that are covered by these regulations, the manual process of compliance checking is costly, time-consuming, and error-prone (Zhang and El-Gohary 2013). In comparison to manual compliance checking, automated compliance checking (ACC) is expected to reduce the time, cost, and errors of the compliance checking process (Zhong et al. 2012, Eastman et al. 2009). To conduct ACC, building information need to be represented in a computer-processable format.

The IFC schema is the most popular building information modeling (BIM) data schema and is becoming the official ISO standard. It was designed as a comprehensive data schema that covers information of all phases of a building and all disciplines of the architectural, engineering, and construction (AEC) industry to support a variety of tasks during a building life cycle including ACC. However, the IFC still lacks

necessary information (i.e., key concepts and relations) that is needed to perform ACC. Different ways to extend the IFC were proposed and/or utilized in previous ACC efforts, such as creating new project parameters (Nguyen and Kim 2011, Sinha et al. 2013), developing new Information Delivery Manuals (IDM) and Model View Definitions (MVDs) (Nawari 2011), and adding new data items (Kasim et al. 2013). Despite the importance of these efforts, previous extension methods mostly extended the IFC in an ad-hoc and subjective manner (i.e., relying on subjective judgements and case-specific developments), and their resulting extended models usually still suffer from missing essential ACC-related information (Niemeijer et al. 2009, Martins and Monteiro 2013). A more generalized and objective method (i.e., relying on rigorous techniques/algorithms for objective judgments that are consistent across different cases) for extending the IFC is, thus, needed.

To address this gap, the authors developed a new method for extending the IFC schema objectively and semi-automatically. The proposed method utilizes semantic natural language processing (NLP) techniques to extract concepts from construction regulatory documents (e.g., building codes) and insert the extracted concepts into the IFC concept hierarchy. NLP aims to enable computers to understand and process natural language text in a human-like manner (Liddy 2001). One key challenge that is faced when developing an IFC extension method is how to relate concepts with no shared terms (e.g., “outdoors” and “outside horizontal clear space”). To address this challenge, in the framework of the proposed IFC extension method, the authors developed a new semantic relation-based concept matching algorithm: the ZESeM (Zhang and El-Gohary Semantic Matching) algorithm. This algorithm aims to match concepts based on their semantic relations. This paper presents the proposed IFC extension method, with a focus on presenting the proposed ZESeM algorithm and the results of testing the algorithm on processing concepts from Chapter 12 of IBC 2006.

2 BACKGROUND

2.1 Natural Language Processing

Natural Language Processing (NLP) is a subdomain of artificial intelligence that aims to enable computers to understand and process natural language text and speech in a human-like manner (Cherpas 1992). NLP has a wide range of applications, such as text classification (Zhou and El-Gohary 2014), information retrieval (Khhokale and Atique 2014), information extraction (Zhang and El-Gohary 2013), text understanding (Karthikeyan and Karthikeyani 2013), and machine translation (Zhao et al. 2014). The analysis used in NLP is categorized into the following levels: (1) morphology, which is the analysis of meaningful components of the words; (2) syntax, which is the analysis of the structural relationship between words; (3) semantics, which is the analysis of meanings of words; (4) pragmatics, which is the analysis of how language is used to accomplish goals; and (5) discourse, which is the analysis on joining of linguistic units larger than utterance and words (Kumar 2011). NLP analysis in each subsequent level is more difficult than its previous level; and the results of the analysis in each subsequent level is more useful than its previous level. State-of-the-Art NLP techniques performed well at the first two levels and is developing fast at the semantic level.

2.2 Semantic Similarity

Semantic Similarity (SS) is the conceptual/meaning distance between two entities such as concepts, words, or documents (Slimani 2013). Semantic similarity plays an important role in information and knowledge processing tasks such as information retrieval (Rodrı´guez and Egenhofer 2003), text clustering (Song et al. 2014), and ontology alignment (Jiang et al. 2014). The measurement of semantic similarity between two entities typically requires established relations (directly or indirectly) between the two entities in an underlying structured knowledge model. Taxonomy and ontology are two types of such knowledge models. A variety of SS computation methods are based on the use of these models.

There are two main types of information that are utilized by SS computation methods: (1) shortest path, and (2) least common consumer. Shortest path is the length of the shortest path (counting nodes or edges) between two entities in a structured knowledge model. Least common consumer is the lowest common superconcept of two entities in a structured knowledge model.

Examples of SS computation methods using shortest path information are Shortest Path Similarity and Leacock-Chodorow Similarity (Table 1). Shortest Path Similarity relies solely on the shortest path information to calculate the SS score between two entities. Leacock-Chodorow Similarity, on the other hand, utilizes the maximum depth of the structured knowledge model in addition to the shortest path information to calculate the SS score between two entities.

Examples of SS computation methods utilizing least common consumer information are Resnik Similarity, Jiang-Conrath Similarity, and Lin Similarity. Resnik Similarity utilizes only the information content of the least common consumer of two entities to calculate the SS score between the two entities. Jiang-Conrath Similarity utilizes the information content of the two entities themselves, in addition to the information content of their least common consumer, to calculate the SS score between the two entities. Lin Similarity takes one step further to use the ratio of the information content of the least common consumer of the two entities to the sum of the information contents of the two entities to calculate the SS score between the two entities (Resnik 1995).

SS provides a measure for semantic-level language analysis, which can be used for many applications that require concept matching. For example, for geographic information service matching, Wang et al. (2013) identified the use of SS as a key enabler of matching, and combined two types of SS measures to enhance the matching efficiency for geographic information services. For ontology alignment, Jiang et al (2014) showed the importance of SS measurement by improving the performance over existing ontology alignment methods through the utilization of a new SS measure. For ontology mapping, Pan et al. (2008) achieved a maximum precision of 80% on ontology mapping using relatedness analysis. Their analysis utilized term-based matching and term co-occurrence statistics. Thus, their method is limited by the corpus used to calculate co-occurrences and may miss concepts that match through general semantic relations because it is difficult for any corpus to capture all semantic relations.

Table 1: Different semantic similarity computation methods and their main information for computation

Existing Semantic Similarity Computation Method	Main Information for Computation
Shortest Path Similarity	shortest path between two entities
Jiang-Conrath Similarity	information content of the two entities themselves, and the information content of their least common consumer
Leacock-Chodorow Similarity	shortest path between two entities, and the maximum depth of the structured model
Resnik Similarity	information content of the least common consumer of two entities
Lin Similarity	ratio of the information content of the least common consumer of two entities to the sum of the information contents of the two entities

2.3 WordNet

WordNet is a lexical database of English developed by Princeton University. In WordNet, nouns, verbs, adjectives, and adverbs are grouped into synsets (sets of cognitive synonyms). Each of the nouns, verbs, adjectives, and adverbs category is organized into a subnet. Four types of semantic and lexical relations are used to link the synsets to one another: synonymy, hyponymy (sub-super or is-a relation), meronymy (part-whole relation), and antonymy (Fellbaum 2005). Synonymy is the semantic relation between different concepts who share the same meaning. For example, “girder” and “beam” share the same meaning of “a horizontal structural component for framework of buildings and other structures.” Hypernymy is a semantic relation where one concept is the hypernym (i.e., superclass) of the other. For example, “structural component” is a hypernym of “beam.” Hyponymy is the opposite of hypernymy, where one concept is the hyponym (i.e., subclass) of the other. For example, “beam” is a hyponym of

"structural component." Meronymy is a semantic relation where one concept is "part of" another concept. For example, "window sill" is a meronym of "window."

Because of the abundant semantic information structurally represented in WordNet, WordNet could be used as the knowledge model for computing SS scores. However, because the basic element in WordNet is a word (or term) rather than concept, WordNet could only be used to compute term-level SS scores using the above-mentioned SS computation methods. Thus, for the remainder of this paper, the above-mentioned SS computation methods are called term-level SS computation methods.

3 PROPOSED METHOD

To extend the IFC schema objectively, the authors developed a four-step method (Figure 1). The method: (1) extracts regulatory concepts from regulatory documents; (2) selects IFC concepts (i.e., concepts in the existing IFC concept hierarchy) that are most relevant to each extracted regulatory concept; (3) classifies the relationship between two concepts for each pair of extracted regulatory concept and selected IFC concept; and (4) constructs the new concept hierarchy by extending the original IFC concept hierarchy with the classified concept pairs. One thing to note in this four-step method is that once a regulatory concept has been added into the IFC concept hierarchy, the regulatory concept becomes an IFC concept.

For the IFC concept selection, a new semantic relation-based concept matching algorithm was developed: the ZESeM algorithm. The rest of the paper presents the ZESeM algorithm and its testing on extending the IFC concept hierarchy with regulatory concepts from IBC.

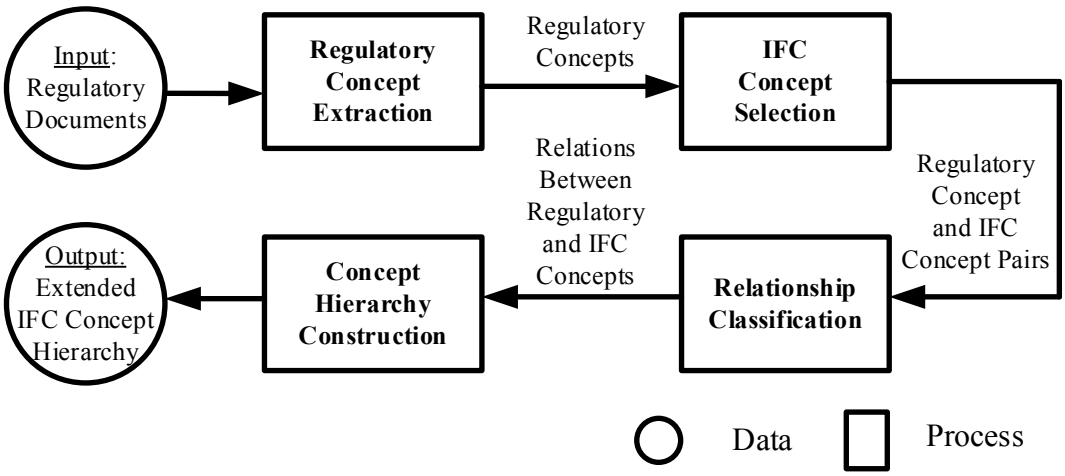

Figure 1: Proposed method to extend the IFC schema objectively

The ZESeM algorithm has three main components (Figure 2): (1) a term-based matching mechanism for finding IFC concepts that share term(s) with an extracted concept; (2) a semantic relation-based matching mechanism for finding related IFC concepts to an extracted concept; and (3) a semantic similarity (SS) scoring function for ranking related IFC concepts according to their relatedness to the extracted concept measured by SS scores.

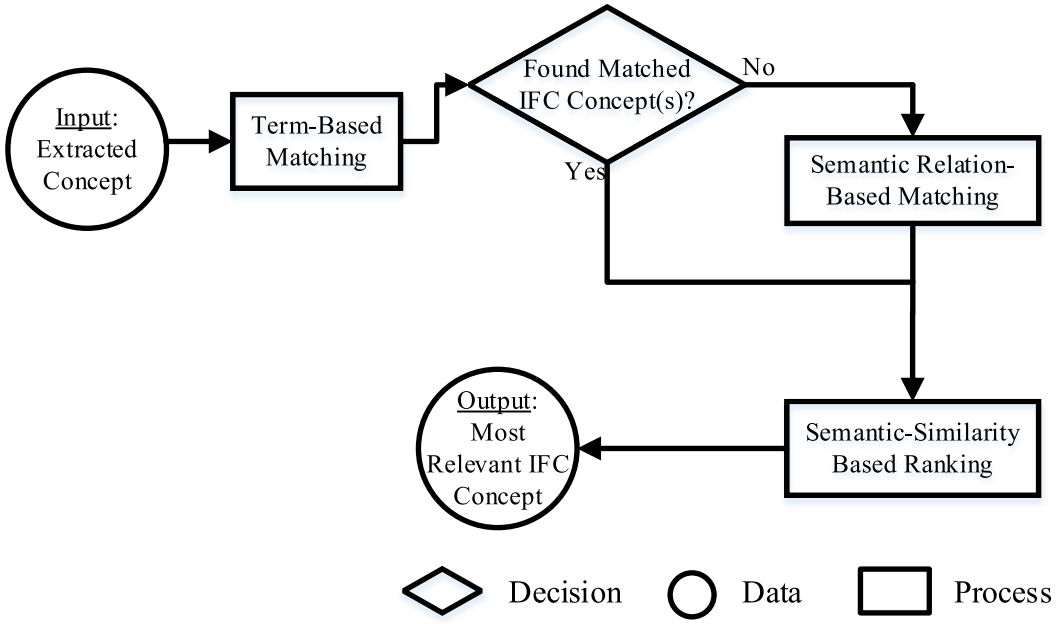

Figure 2: The proposed ZESeM (Zhang and El-Gohary Semantic Matching) algorithm

Depending on the part-of-speech pattern of an extracted concept, the first term and last term of this concept are used to search for matched IFC concepts in both term-based and semantic relation-based matching mechanisms. If the extracted concept only has one term or the first term of the extracted concept is not a noun (i.e., singular or mass noun, plural noun, gerund, or proper noun), then only the last term of the extracted concept is used to search for matched IFC concepts. If the extracted concept has multiple terms and the first term of the extracted concept is a noun, then the first term of the extracted concept is used to search for matched IFC concepts, in addition to the last term. The matching term (first term or last term of the extracted concept) is compared with each term of an IFC concept, and the IFC concept is preliminarily selected if it includes at least one occurrence of the matching term of the extracted concept. In both matching mechanisms, stemming is used to ensure that matches in different forms of a word are not missed. For example, through stemming, “foot” could be matched with “feet”, and “reinforcing” could be matched with “reinforced”. In the term-based matching mechanism, the matching is pure term-based (i.e., string match) with stemming applied to both terms. In the semantic relation-based matching mechanism, the matching is based on three types of semantic relations: hypernymy, hyponymy, and synonymy (Fellbaum 2005).

For each extracted concept, term-based matching is applied first. If no matches are found using term-based direct matching, semantic relation-based matching is applied. As such, after the matching is conducted, there are three possible consequences: no matched IFC concept found, one matched IFC concept found, and more than one matched IFC concept found. In the first case, the extracted concept will be abandoned. In the second case, the matched IFC concept is selected. In the third case semantic-similarity based ranking is applied to find the highest ranked IFC concept; the highest ranked IFC concept is then selected.

Equation 1 shows the proposed SS scoring function that was used in the third case. The meaning of each parameter in the function is explained as follows:

1. SS_{RF} is the SS score between extracted concept R and IFC concept F.

2. SS_{RmFk} is the term-level SS score between the matching term m in the extracted concept R and the k_{th} term in the IFC concept F. SS_{RmFk} is calculated utilizing existing term-level SS score computation methods such as Shortest Path Similarity and Leacock-Chodorow Similarity.

3. 2k/(n(n+1)) is a term pair discount factor in which k is the ordinal number for the term Fk in IFC concept F and n is the length of F measured in number of terms.

4. 1/(n) is the final discount factor which linearly discounts the summation using the length of the IFC concept.

[1] $$SS_{RF} = \frac{1}{n}\sum_{k=1}^{n}\frac{2k}{n(n+1)}SS_{RmFk}$$

The term pair discount factor 2k/n(n+1) is based on the heuristic that in a multi-term concept, the contribution of each term's carried meaning to the meaning of the whole concept decreases from right to left. The final discount factor 1/(n) is based on the heuristic that the length of a concept name is related to its level in a concept hierarchy. The shorter the length of a concept name, the more general the concept is; and thus the higher its level in a concept hierarchy.

4 EXPERIMENTAL EVALUATION

The ZESeM algorithm was tested on processing extracted concepts from Chapter 12 of IBC 2006. IBC was selected because of its prevailing adoption in the United States (adopted by 46 states). Chapter 12 of IBC 2006 was then randomly selected. The longest span for each noun phrase in Chapter 12 of IBC 2006 was manually recognized and extracted as a concept. For example, concepts in the list L1 were recognized and extracted from Sentence S1. In total, 368 concepts were extracted. WordNet (Fellbaum 2005) was used to define the semantic relations between terms. Five term-level SS computation methods were tested in the ZESeM algorithm for comparison: Shortest Path Similarity, Leacock-Chodorow Similarity, Resnik Similarity, Jiang-Conrath Similarity, and Lin Similarity (Resnik 1995). The term-based matching and semantic relation-based matching mechanisms were evaluated separately. The evaluation was conducted using the measure of adoption rate, which was defined as the number of found IFC concepts that were adopted divided by the total number of found IFC concepts. The evaluation was conducted using a gold standard.

- S1: "The minimum net area of ventilation openings shall not be less than 1 square foot for each 150 square feet of crawl-space area."

- L1: ['minimum_net_area', 'ventilation_openings', 'square_foot', 'square_feet', 'crawl-space_area']

The ZESeM algorithm was implemented using Python programming language (v.2.7.3). The "re" (regular expression) module in python was utilized to implement the matching mechanisms. The hypernymy, hyponymy, and synonymy relations in WordNet were utilized through the NLTK (Natural Language Toolkit) WordNet interface in python. The Porter Stemmer (Porter 1980) was utilized to implement stemming.

5 EXPERIMENTAL RESULTS

The experimental results for term-based matching and semantic relation-based matching are summarized in Table 2. The adoption rate for different term-level SS computation methods ranged from 80.7% to 87.1%. For term-based matching, the highest adoption rate exceeds the lowest adoption rate by 5.1%. For semantic relation-based matching, the highest adoption rate exceeds the lowest adoption rate by 2.2%. Table 3 shows some randomly selected example concepts that were extracted and matched using the different term-level semantic similarity computation methods, for term-based matching and semantic relation-based matching. The matched IFC concepts that were not adopted are shown in italics.

Table 2: Performances of different term-level semantic similarity computation methods on term-based matching and semantic relation-based matching

Matching Mechanism	Existing Semantic Similarity Algorithm	Number of IFC Concepts Found	Number of IFC Concepts Adopted	Adoption Rate
Term-based	Shortest Path Similarity	286	249	**87.1%**
	Resnik Similarity	286	246	86.0%
	Lin Similarity	286	246	86.0%
	Jiang-Conrath Similarity	286	244	85.3%
	Leacock-Chodorow Similarity	286	237	82.9%
Semantic relation-based	Shortest Path Similarity	114	94	**82.5%**
	Resnik Similarity	114	93	81.6%
	Lin Similarity	114	93	81.6%
	Jiang-Conrath Similarity	114	92	80.7%
	Leacock-Chodorow Similarity	114	93	81.6%
Overall	Shortest Path Similarity	400	343	**85.8%**

6 DISCUSSION

Shortest Path Similarity achieved the best adoption rate for both matching mechanisms. Shortest Path Similarity only utilizes the shortest path between two concepts in a concept hierarchy for SS score computation, whereas Leacock-Chodorow Similarity utilizes both the shortest path between two concepts and the maximum depth of concepts in the hierarchy and other term-level SS computation methods utilize information content of the least common consumer (i.e., the lowest concept in the concept hierarchy that is the superclass of the two concepts). The higher performance of Shortest Path Similarity over Leacock-Chodorow Similarity indicates that the additional information of concept depth did not help in concept matching. The higher performance of Shortest Path Similarity over Jiang-Conrath Similarity, Resnik Similarity, and Lin Similarity indicates the advantage of the shortest path over the information content of the least common consumer in measuring semantic similarity for concept matching. Thus, based on the comparative experimental results of different term-level SS computation methods, the Shortest Path Similarity was selected as the term-level SS computation method in the ZESeM algorithm.

Both term-based matching and semantic relation-based matching achieved good performance of over 80% adoption rate. The examples in Table 3 show the effectiveness of the ZESeM algorithm. The matches between “interior spaces” and “space” and between “square foot” and “feet” show the effectiveness of term-based matching, including the use of stemming. The matches between “floor joists” and “beam”, between “enclosed attics” and “additional story”, and between “cornice vents” and “openings” show the effectiveness of semantic relation-based matching in leveraging the semantic relations between terms to find concept matches that would have otherwise been missed (if only utilizing term-based matching, without the use of semantic information, because those concept pairs do not share any terms).

The processing using term-based matching has larger variation in performance than that using semantic relation-based matching (adoption rate relative difference of 5.1% compared to 2.2%) among the term-level semantic similarity computation methods. Because term-based matching can easily find more candidate IFC concepts to match than semantic relation-based matching, more candidate IFC concepts lead to a higher probability of variability in performance. However, semantic relation-based matching finds semantically related concepts that cannot be found using term-based matching. Thus the combination of semantic relation-based matching and term-based matching finds more matched concepts than using either one of them.

Table 3: Examples of extracted concepts and matched IFC concepts using different term-level semantic similarity computation methods

Extracted Concept	Matching Mechanism	Existing Semantic Similarity Computation Method	Matched IFC Concept
Ventilators	Term-based	Shortest Path Similarity	*Cross ventilation*
		Jiang-Conrath Similarity	*Ventilation*
		Leacock-Chodorow Similarity	*Ventilation*
		Resnik Similarity	*Ventilating opening*
		Lin Similarity	*Ventilating opening*
Interior spaces	Term-based	Shortest Path Similarity	Space
		Jiang-Conrath Similarity	Space
		Leacock-Chodorow Similarity	Space
		Resnik Similarity	Space
		Lin Similarity	Space
Square foot	Term-based	Shortest Path Similarity	Feet
		Jiang-Conrath Similarity	Feet
		Leacock-Chodorow Similarity	*Footing*
		Resnik Similarity	Feet
		Lin Similarity	Feet
Floor joists	Semantic relation-based	Shortest Path Similarity	Beam
		Jiang-Conrath Similarity	Beam
		Leacock-Chodorow Similarity	Beam
		Resnik Similarity	Beam
		Lin Similarity	Beam
Enclosed attics	Semantic relation-based	Shortest Path Similarity	Additional story
		Jiang-Conrath Similarity	Additional story
		Leacock-Chodorow Similarity	Additional story
		Resnik Similarity	Additional story
		Lin Similarity	Additional story
Cornice vents	Semantic relation-based	Shortest Path Similarity	Openings
		Jiang-Conrath Similarity	Openings
		Leacock-Chodorow Similarity	Openings
		Resnik Similarity	Openings
		Lin Similarity	Openings

7 LIMITATIONS AND FUTURE WORK

One limitation of this study is that only semantic relation-based matching on unigram (single terms) was used for finding semantically related IFC concepts to an extracted concept. While the combinatorial nature of term meanings [i.e., the meanings of single terms (e.g., “exterior” and “door”) in a concept name are combined to form the overall meaning of the whole concept (e.g., “exterior door”)] renders this unigram method effective, there may be cases where semantic relation-based matching on bigram (pairs of terms) or multigram (groups of three or more terms) could be effective. As such, in future work, the authors plan to extend the semantic relation-based matching mechanism to incorporate semantic relations between bigram and multigram to test whether such bigram or mutligram considerations could further improve the performance of concept matching.

8 CONCLUSION

This paper presents a new natural language processing (NLP)-based method for extending the IFC schema objectively. As part of the proposed IFC extension method, a new semantic relation-based concept matching algorithm, called ZESeM algorithm, was developed. The ZESeM algorithm utilizes both term-based matching and sematic relation-based matching to find matching IFC concepts for extracted regulatory concepts. A new function was proposed to compute concept-level semantic similarity (SS) scores between concepts based on their term-level SS scores. The proposed algorithm was tested on finding matching IFC concepts for extracted concepts from Chapter 12 of IBC 2006. The experimental results verify the effectiveness of the proposed concept-level SS function and the proposed ZESeM algorithm in concept matching. An 85.8% adoption rate was achieved. For term-level SS computation, Shortest Path Similarity showed the best performance. One limitation of the proposed method is that only unigram (single terms) semantic relation-based matching was used, while bigram or multigram semantic relation-based matching may further improve the matching performance. In their future work, the authors plan to extend the semantic relation-based matching mechanism to further test the effectiveness of bigram and multigram semantic relation-based matching.

Acknowledgements

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 authors and do not necessarily reflect the views of NSF.

References

Cherpas, C. 1992. Natural language processing, pragmatics, and verbal behavior. *The Analysis of Verbal Behavior*, **10**(1992): 135–147.

Eastman, C., Lee, J., Jeong, Y., and Lee, J. 2009. Automatic rule-based checking of building designs. *Automation in Construction*, **18**(8): 1011–1033.

Fellbaum, C. 2005. WordNet and wordnets. *Encyclopedia of Language and Linguistics, Second Edition*, Elsevier Limited, Oxford, South East England, UK, 665-670.

Jiang, Y., Wang, X, and Zheng, H. 2014. A semantic similarity measure based on information distance for ontology alignment. *Information Sciences*, **278**(2014): 76-87.

Karthikeyan, K., and Karthikeyani, D.R. 2013. Understanding text using anaphora resolution. *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME)*, IEEE, Salem, Tamil Nadu, India, 346-350.

Kasim, T., Li, H., Rezgui, Y., and Beach, T. 2013. Automated sustainability compliance checking process: proof of concept. *CONVR 2013, 13th International Conference on Construction Applications of Virtual Reality*, Teesside University, London, UK, 11-21.

Khhokale, R.S., and Atique, M. 2014. Intelligent interface for web information retrieval with document understanding. *Human-Computer Interaction, Part III, HCII 2014, LNCS 8512*, Springer International Publishing Switzerland, Cham, Switzerland, 21-31.
Kumar, E. 2011. *Natural language processing*. I.K. International Publishing House Pvt. Ltd., New Delhi, India.
Liddy, E.D. 2001. Natural language processing. *Encyclopedia of Library and Information Science, 2nd Ed*. Marcel Decker, Inc., New York, NY, USA.
Martins, J.P., and Monteiro, A. 2013. LicA: A BIM based automated code-checking application for water distribution systems. *Automation in Construction*, **29**(2013): 12–23.
Nawari, N.O. 2011. Automating codes conformance in structural domain. *International Workshop on Computing in Civil Engineering 2011*, ASCE, Miami, Florida, USA, 569-577.
Nguyen, T., Kim, J. 2011. Building code compliance checking using BIM technology. *2011 Winter Simulation Conference (WSC)*, IEEE, Phoenix, AZ, USA, *3395-3400.*
Niemeijer, R.A., Vries, B.D., and Beetz, J. 2009. Check-mate: automatic constraint checking of IFC models. *Managing IT in Construction/Managing Construction for Tomorrow*, CRC Press, London, UK, 479-486.
Pan, J., Cheng, C.J., Lau, G.T., and Law, K.H. 2008. Utilizing statistical semantic similarity techniques for ontology mapping - with applications to AEC standard models. *Tsinghua Science and Technology*, **13**(S1): 217-222.
Porter, M. 1980. An algorithm for suffix stripping. *Program (Automated Library and Information Systems)*, **14**(3): 130-137.
Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *14th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, **1**:448-453.
Rodrı´guez, M.A., and Egenhofer, M.J. 2003. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering,* **15**(2): 442-456.
Slimani, T. 2013. Description and evaluation of semantic similarity measures approaches. *International Journal of Computer Applications*, **80**(10): 25-33.
Sinha, S., Sawhney, A., Borrmann, A., and Ritter, F. 2013. Extracting information from building information models for energy code compliance of building envelope. *COBRA 2013 Conference*, International Council for Research and Innovation in Building and Construction (CIB), New Delhi, India.
Song, W., Liang, J.Z., and Park, S.C. 2014. Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering. *Information Sciences*, **273**(2014): 156-170.
Wang, S., Yu, H., and Su, X. 2013. A semantic similarity algorithm for geographic information service matching. *Applied Mechanics and Materials*, **405-408**(2013), 3070-3074.
Zhang, J., and El-Gohary, N.M. 2013. Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. *Journal of Computing in Civil Engineering*, 10.1061/(ASCE)CP.1943-5487.0000346, 04015014.
Zhong, B. T., Ding, L. Y., Luo, H. B., Zhou, Y., Hu, Y. Z., and Hu, H. M. 2012. Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking. *Automation in Construction*, **28**(2012): 58–70.
Zhou, P., and El-Gohary, N. 2014. Ontology-based multi-label text classification for enhanced information retrieval for supporting automated environmental compliance checking. *2014 International Conference on Computing in Civil and Building Engineering,* ASCE, Orlando, Florida, USA, 2238-2245.
Zhao, Y., Huang, S., Chen, H., and Chen, J. 2014. An investigation on statistical machine translation with neural language models. *CCL and NLP-NABD 2014, LNAI 8801*, Springer International Publishing Switzerland, Cham, Switzerland, 175-186.