Automated generation of highway model view using NLP

Tuyen Le [1], David Jeong[2]

(To be submitted to the Journal of Computing in Civil Engineering)

ABSTRACT

Open data standards (e.g. LandXML, TransXML) have been widely recognized as the solution to the interoperability issue in exchanging digital data in the transportation sector. Since these schema structure rich sets of data containers covering a wide range of fields and phases, model view definitions (MVDs) which define subsets of schema in accordance with the exchange requirements for specific scenarios are required. The process of MVD development is time consuming as the developer has to manually search for the entities and attributes names that semantically match to the data exchange requirements. This paper presents a framework that can automatize the process of mapping data labels based on their semantic similarity. The framework employs an unsupervised machine learning to learn the semantic relatedness between technical concepts from an unlabeled highway related corpora as the input data. The input corpus includes 10 million words mainly collected from roadway design guidelines across U.S. States, New Zealand and Australia. The model will be tested and evaluated by comparing the mapping results performed by computer and experts.

Keywords: Ontology, NLP, model view, interoperability, data exchange, highway

INTRODUCTION

Neutral data standards have been widely accepted as the solution to the interoperability issue in the construction industry. Several open standards have been proposed, ranging from solely relying on syntactics using Express Modeling Language such as Industry Foundation

[1]Ph.D. Student, Department of Civil, Construction and Environmental Engineering, Iowa State University. Ames, IA50011. E-mail: ttle@iastate.edu.

[2]Associate Professor, Department of Civil, Construction and Environmental Engineering, Iowa State University. Ames, IA 50011. E-mail: djeong@iastate.edu.

Classes (IFC) (buildingSMART 2015) or LandXML (landxml.org 2015) to semantics-rich ontologies such as e-COGNOS (Lima et al. 2005). These standardized data models consist of rich sets of data elements covering various business processes and disciplines. However, since a specific data exchange scenario needs only a subset of data, hence neutral data standards alone are insufficient to facilitate seamless digital data exchange among project stakeholders (Froese 2003; East et al. 2012). As querying data on those data schema which are large and complicated the end user is required to have considerable programming skills and properly understand the structure and the meaning of each entity or attribute included in the source data schema. Data driven decision making based on a wrong extracted dataset would likely lead to a wrong decision. Therefore, there is a need for the formal definitions of schema subsets determining the right data for specific transactions. The availability of these model views will underpin the extraction of data from complicated sets of data generated from the AEC industry.

To address the above demand, a considerable amount of research efforts has been made in both the building and transportation sectors with the same ambition to define subsets of data for various business processes. One of these efforts is the Construction to Operation Building Information Exchange (Cobie) project (East 2007) which is now becomes a part of a variety of national standards and guidelines for projects using Building Information Modeling (BIM), for instance UK COBie 2.4 (Nisbet 2012), National BIM Standard-United States Version 3 (NBIMS-US) (National Institue of Building Sciences (NIBS) 2015), GSA-BIM Guide (U.S. General Services Administration (GSA) 2011). This research identified IFC data elements that are generated in the design and construction phases required to be transferred to the asset management phase. The civil sector also is going on this trend with several model views of the Landxml schema has been being defined. The examples of these include the InfraModel project carried by the Technical Research Center of Finland aims to specify subsets of LandXML schema for several transportation projects and this specification has become the Finish national application specification (inframodel.fi 2014).

Even though a considerable number of research have been made, but these are still limited to a large demand from the industry. This is because the current method for developing model view definition is based on a manual basic which is time consuming (Venugopal et al. 2012; Eastman 2012; Hu 2014). The business processes are dynamic and tend to change over time. To adapt to the changes from industry practices, these model view are required to be tailored. Therefore, there is a need to change the current practice of model view definition from the ad-hoc approach to a more rigorous methodology (Venugopal et al. 2012).

To address the above issue, this paper aims to propose a novel method that allows for automated generation of model views from a highway data schema. The proposed framework are consisted of the following two key modules. The first module employs Natural Language Processing (NLP) techniques to read domain-specific guidelines and extract data requirements. In the second module, a matching algorithm was deployed to match the extracted data requirement entities to the AEC (architect, engineering and construction) data schema that describing the data generated in the up stream phases. The core component of the matching algorithm is the machine learning based disambiguator which was trained to automatically understand senses of technical terms and match terms based on their semantics instead of their pure labels. The training data set is a highway text corpus which was collected from a number of documents from multiple sources such as Wikipedia, textbooks, federal and state highway engineering manuals, research reports for across project phases including design, construction and asset management. The java package of the developed algorithm is stored at https://github.com/tuyenbk/mvdgenerator.

LITERATURE REVIEW

Natural Language Processing

NLP is a collection of techniques that can analyze and extract information from natural language like text and speech. The major applications of NLP include translation, information extraction, opinion mining (Cambria and White 2014). These applications are supported by a combination of several techniques such as Named Entity Recognition (NER),

Part-of-Speech (POS) tagging (Toutanova et al. 2003; Cunningham et al. 2002), tokenization (or word segmentation) (Webster and Kit 1992; Zhao and Kit 2011), relation extraction, sentence parsing, word sense disambiguation (Lesk 1986; Yarowsky 1995; Navigli 2009), etc. NLP methods can be classified into two main groups: (1) rule-based and (2) machine-learning (ML) based methods. Since the early group, rule-based NLP, was based solely on hand-coded rules, these systems are not able to cover the complicated set of human grammatical system (Marcus 1995) and, therefore, do not perform well. The current trend in NLP research is the shift from rule based analysis to statistical ML based methods (Cambria and White 2014). ML models are able to learn patterns from training examples to predict the output, hence they are independent to languages, linguistic grammars and consequently reduce human resources cost (Costa-Jussa et al. 2012).

Word Sense Disambiguation (WSD) is one of the main NLP related research topics. WSD aims to measure the similarity/relatedness between semantic units (words, sentence, concepts, etc.) (Harispe et al. 2015). Semantic measurements can be classified into two groups that are corpus based and context based approaches (distributional measures and knowledge based measures) (Harispe et al. 2013). The corpus based approach relies on the predefined lexicons which provide structured vocabularies. One of the popular measurements of similarity between two concepts is measure the distance between concepts in the lexical hierarchy. The shortest path between concepts presents the relatedness between them. The second approach is based on the context of two concepts. This approach eliminates the reliance on the availability of lexicon which may mot available. In this approach, two concepts are considered to be similar if they have the same context (surrounding concepts) in a corpus of text.

Digital data query method

Conventional method are to develop a language for querying information from dataset stored in neutral data format. these research aims to provide a query language specific for the querying data BIM data. (**?**) presented a review of existing query approach ranging

from SQL-based, Relational-based, Object orientation based, xml-based . Object-relational database approach to query IFC model (?) that outperformed tradition relational databases servers. Spatial query language , semantic and spatial conditions used for bim query (?). graph showing adjacent and accessiblity relations between spaces, that can be used to build a look-up mechanism in BIM instance query and developed a BIM query language called QL4BIM Spatio-semantic query language (?) space related information; graph-based retrieval of BIM (?), search strategy; graph-based for topological querying BIM elements (?).

efforts have been made to simplify which allow user easier way to interact with complicated sets of digital information. the above research aims to enhance data search strategy and machine-redable language which still requires user to acknowledge new data schema and query language. these aims to explicit relations between entities which are implictly represented in the neutral data schema. this allow reduce burden on user in acknowledging the data schema. ontology-driven construction information retrieval (?) for tunnel projects; ontology partial BIM model extraction (?) for generic method for building projects; speicific for construction information, ontology-based extraction of construction information (?). this approach allow for reasoning integration. query over linked life cycle data spaces (?).

a further step to enhance the process of data extraction is to provide tools that requirements less effort in studying data structure and provide visual way for the end user to extract data. no-schema algorithm for extracting IFC instances without using ifc schema or model view definition (?). visual BIM query (?), visual bim query language, predefined library, basically a visualization of query code. query data with mvd as input data, (?). online BIM resources retrieval system (?).

in summary, lack of human-computer interaction that enable computer understand and interpret the end user interest.

model view definition development process

IFD-A standard on BIM data mapping

Word sense ambiguity is seen as one of the major interoperability issues in data integration or exchange between isolated sources. This section present current efforts in handling the issue of ambiguity in data exchange between domains, languages by buildingSMART.

IFD Library (International Framework for Dictionaries library) is one of the component of the pillar which is visioned by buildingSMART as the solution to the interoperability issue in the construction industry. IFD (ISO 12006-3) is developed by the IFD Library Group which is a part of buildingSMART International, including buildingSMART Norway, Construction Specifications Canada (CSC), U.S. Construction Specification Institute (CSI), and STABU Foundation. IFD is an EXPRESS based model that supports the creation of multilingual dictionaries or ontologies and offers a standardized method for semantics mapping among concepts in IFC models (Bjï£¡rkhaug and Bell n.d.). The IFD mapping is based on semantics rather than word similarity. IFD assigns each concept/vocabulary in the dictionary/library a Global Unique Identifier (GUI) or Universal Unique Identifier (UUID) and this identical value is used through the communication process instead of the IFC entity name (Bjï£¡rkhaug and Bell n.d.). buildingSMART Data Dictionary is an example of dictionary developed based upon ISO 12006-3. The dictionary includes building vocabularies supplement for the development of IFC and data exchange process. IFD is designed to make two bilingual words are semantically mapped. concepts are distinguished by a unique id. two words/synonyms (in bilingual languages) will be assigned the same id value. the exchange are based on the matching of id which represent concepts instead of words.

Other academic research on semantic mapping

In the construction industry, research efforts are currently focusing on standardizing the data structure format, there are few research have been done to deal with the issue of sense ambiguity. (Zhang and El-Gohary 2015) proposed an algorithm call ZESeM for determining the semantic similarity between two concepts. The algorithm includes two

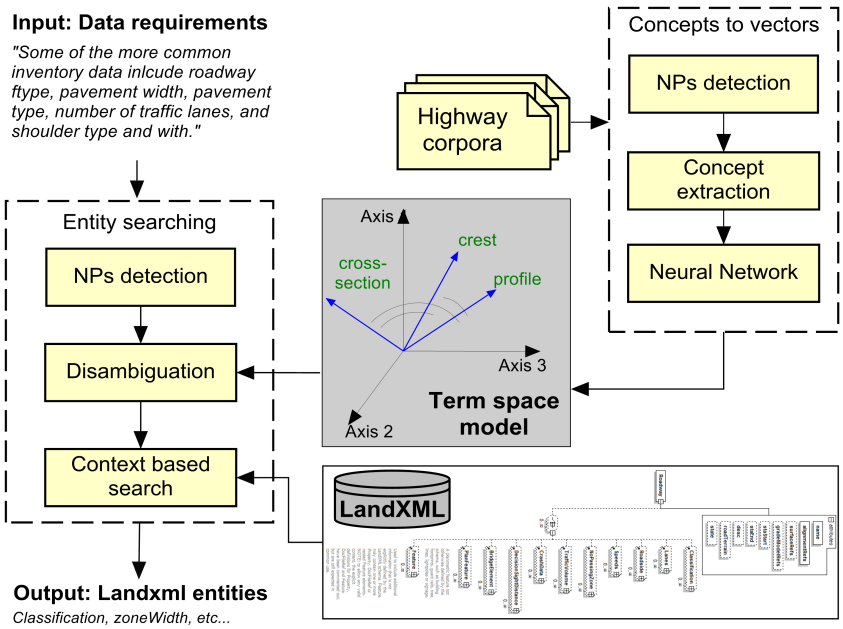

FIG. 1: Overall architecture for automated generation of MDV

sequential steps that are a term-based matching and a semantic relation based matching. One problem with this algorithm is the lack of the integration of syntax and context features, hence disambiguation still remains in the case in which the same word form is used for different senses. The algorithm accepts results from the label based matching step. But those matching may have different semantics. Another semantic mapping framework (Lin et al. 2015) which was based on IFD to map user's requirements to the IFC schema for data extraction. These previous research were both based the hand-made electronic vocabularies. ZESeM relied on Wordnet, and Lin' model employs vocabulary set for building industry IFD. Wordnet is generic lexicon, it lacks of technical terms for the construction industry. IFD is more building specific industry but it still just cover a small domain in the industry. Since manually constructed e-dictionary is time consuming and therefore still very limited to large demand and dynamic world. There are still many domain/areas are not covered in these dictionaries.

OVERALL ARCHITECTURE

Figure 1 presents the proposed framework for automated generation of model views. The

framework is consisted of two components that are: (1) a highway term space model, and (2) a semantic searching algorithm. The first module aims to extract highway related technical concepts from the highway corpora and transform concepts into vectors representing their meanings by employing a neural network (NN) model and a set of NLP techniques. Using this concept vector space, synonyms or associated concepts can be determined based on the distance or angle between vectors. The purpose of the second module is to semantically search for LandXML entities and attributes based on the natural language data requirement inputs. In order to achieve this objective, NLP techniques firstly are applied on the natural language data requirements to extract keywords that representing what types of data needed to be transferred to the data receiver. A proposed algorithm then is utilized to disambiguate the meaning of extracted required data keywords based on the term space model and search for equivalent entities or attributes included in the LandXML data schema. The following sections respectively presents the process of building the highway term space model and the searching algorithm along with details on which methods/tools utilized

HIGHWAY TERM SPACE MODEL

The ultimate goal of this module is to building a model that can support the disambiguation task. For disambiguation, there are several methods including thesaurus based, ontology based and distributional method. The first two methods required a full lexicon or ontology including concepts description for all aspects/disciplines in the highway industry. These methods would be ideal for the disambiguation task if domain related thesauruses are available. However, since building up those dictionaries required a huge amount of empirical work, they are still limited. Wordnet (Miller 1995) which is one of the largest lexicons available containing 117,000 synsets, but it is generic and is not suitable for the highway domain. For this reason, this research employs the distribution method which is based on an unsupervised machine learning method to train unlabeled data and learn the meaning of words by analyzing the context of words.

Highway vector space model (H-VSM) is one of the key results of this research. The

skip-gram model, proposed by Mikolov et al. (2013), which is a neural network (NN) model was employed to develop the H-VSM. The NN model was designed to predict the context words of a given word included in the input corpora. The sub-sections below present the detailed procedure followed to collect input data which is the highway domain corpora and utilize the skip-gram model to convert highway technical concepts to vectors.

Data collection

Highway corpus was collected from a number of documents from multiple sources including textbooks, and highway engineering manuals from federal Department of Transportation and from 18 state DOTs. The focus of highway corpora in this this research were on three project phases including design, construction and asset management. Since technical guidance documents in the engineering field contains a variety of text formats such as text, tables, equations. The order of words in tables and equations are not structured in the sentence format, their words are not suitable for the training process. Hence, these tables were removed from the text corpora. The result of the data collection was a highway corpora consisting of 10 millions words. This data set was utilized to extract highway related technical terms which were then trained and converted into vectors.

Concept extraction

The first step in building the H-VSM is the identifying of highway related technical concepts. In order to achieve this task, a set of natural language processing techniques including tokenization, part of speech tagging were utilized to identify the POS tag for each word in the highway corpora. The OpenNLP library was used to perform this task. The linguistic process, as illustrated in the Figure 2, includes the following steps:

Word Tokenization

In this step, text was broken down into individual units (called tokens).

POS tagging

The purpose of this step is to determine the part of speech tag for each token.

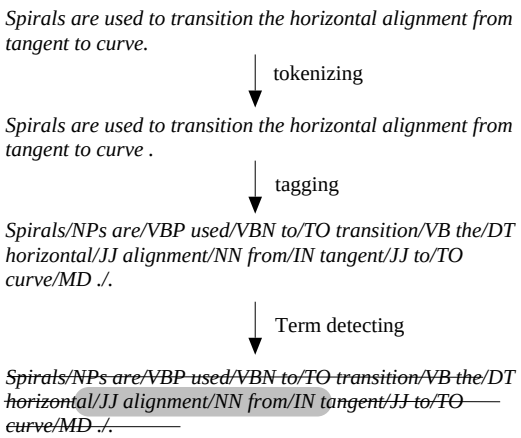

FIG. 2: Linguistic processing procedure to detect technical terms

Noun phrase detection

For this task, any phrases having either the following patterns (A|N)1*N1 Prep(of) (A|N)2*N2 or (A|N)1*N1(A|N)2N2 are categorized as noun phrases and are good candidates for technical terms. In the patterns above, A is adjective and N is noun.

Termhood measurement and concept extraction

The C-Value algorithm, proposed by (Frantzi et al. 2000) was employed to extract technical concepts. The idea of the C-value method is to firstly extract noun phrases which are composed of adjective and nouns, and then measure the how frequent they occur in the corpus. Equation below presents the C-value measurement of termhood.

$$C-value(a) = \begin{cases} log_2|a|.f(a), & \text{if a is not nested} \\ log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)), & \text{otherwise} \end{cases} \quad (1)$$

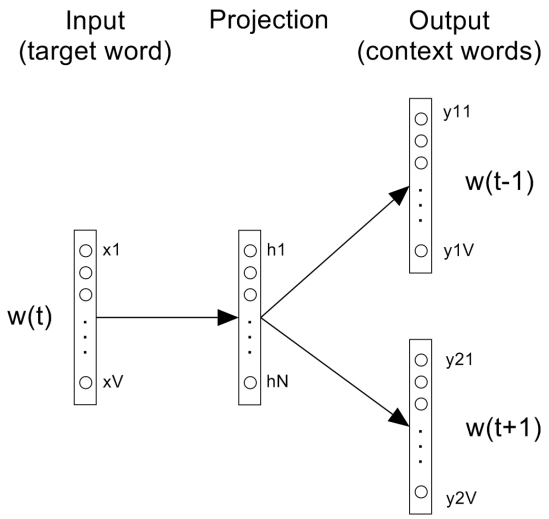

FIG. 3: Skip-gram model

where:

a is a candidate noun phrase

f is the frequency of a in the corpus

Ta is the set of extracted noun phrases that contains a

P(Ta) is the number of these candidate terms.

Data training

This research employed the skip-gram neural network training model which was developed by (Mikolov et al. 2013) to train the highway corpus. The distributional hypothesis is the fundamental theory of this method. The distributional hypothesis says that two words have the same meanings if they occur in the same contexts (Harispe et al. 2015). For example, "apple" is more similar to "banana" than coffee since "apple" and "banana" are both co-occur with the word "eat", while "coffee" more frequently co-occur with the word "drink". Figure 3 presents the architecture of the skip-gram model. The model includes 3 layers including input, hidden and output layer. The input data is the target word and the output data are the context words surrounding the target word. Each word in the vocabulary set (with the

TABLE 1: Skip-gram model parameters

Parameter	Value
Hidden size	300
Window size	15

size V) is encoded into one-hot vector in which only the value at the word index is equal to one and other positions are zero. In order to train the model, the java library word2vec, developed by Google was used. The parameters of the model, as presented in Table 1, were selected based on the suggestions in the literature.

CONTEXT BASED SEARCHING ALGORITHM

This section presents the proposed algorithm (see Figure 4) for semantically searching for equivalent entities/attributes in Landxml schema. The algorithm is a three-stage procedure. In the the first stage, a list of synonym concepts that have the similar meanings to the keyword input are generated by utilizing the vector space model developed above. Each synonym concept is defined by their attributes. A string based searching is then applied to find the entity in the Landxml schema that has the most similar name for each synonym in the list. Any synonym that match to at least one Landxml entity in this phase is considered as potential candidate. These candidates are then tested in the final phase. Concepts which fail the testing phase are removed from the candidate list. The test is to examine if the synonym and it's matched entity share the same attributes. If two objects are similar, they would share common attributes. The final matches are ranked by the similarity score.

EVALUATION

An evaluation experiment was conducted to evaluate the context-aware searching algorithm. In this experiment, a graduate student was asked to read a randomly selected document which contains data requirements for a specific data transaction. The duty of the student was to manually identify a set of Landxml entities that would fulfill the data requirements. Meanwhile, a prototype built upon the developed algorithm was applied to

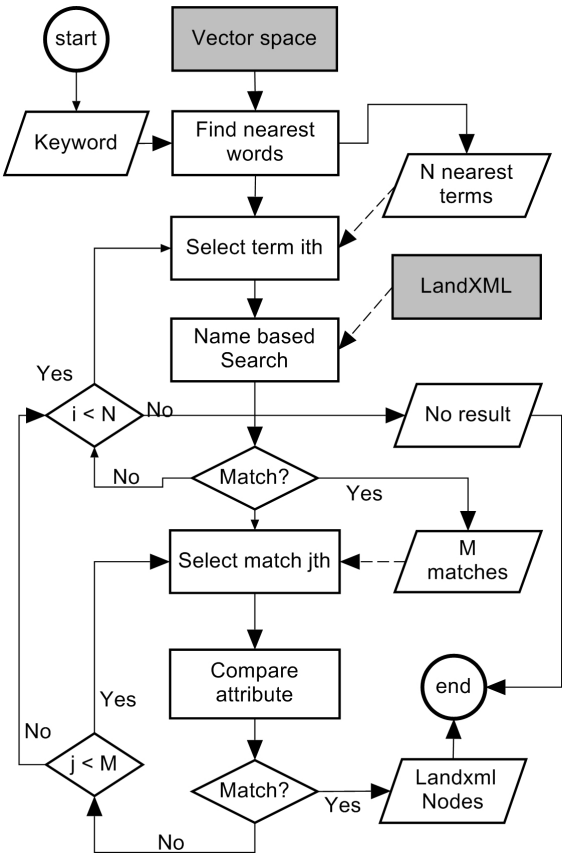

FIG. 4: Entity searching algorithm

automatically generate the subset of data from the same document as the student used. The results from the two methods were used to calculate the accuracy of the searching algorithm.

$$Recall = \frac{\text{number of correctly matched concepts}}{\text{total concepts}} \tag{2}$$

$$Precision = \frac{\text{number of correctly matched concepts}}{\text{total matched concepts}} \tag{3}$$

$$F - measure = \frac{2.Precision.Recall}{Precision + Recall} \tag{4}$$

Table 2 shows the evaluation result. As presented in the table, the system shows a 90 percent precision. However, the recall is relatively low accuracy, this is possibly due to the the training data size. Since the searching algorithm accuracy is highly rely on the a capacity

of finding synonyms which is based on the vector space model. This model currently is based on the data training set consisting of only 10 million words. In order to enhance the accuracy, the data training set needs to be extended. Future research will be conducted to extend the training data set.

CONCLUSIONS

Digital data has been widely generated through the project life cycle. However, the data collected and generated in previous stages are no reusable in the downstream phases. This issue is due to the interoperability when digital data from one partner is not readable or correctly understandable by the data receiver. This research develops an framework that semantically searches for desired data from the transferred data file. The framework is composed of two components including (1) a terms space model which represents highway related concepts extracted from the highway corpora in vectors and (2) a context based searching algorithm that can search for entities in the Landxml schema based on their similarity of attributes instead of string based similarity.

The framework has been evaluated by testing on a randomly selected set of input data. The result shows the accuracy of over 80 percent. The accuracy is low due to the size of the training data. Future research will be conducted to increase the data size.

This method is broad and can be applied to other business processes such as green building checking, environment checking, etc. The method is expected to significantly improve the existing ad-hoc method of model view definition development and in return leads the the removal of this bottle neck which is restricting the seamless data integration and exchange across phases of a highway construction project.

REFERENCES

TABLE 2: Evaluation result

Tested concepts	Matched concepts	Correct	Recall(%)	Precision (%)	F (%)
23	21	19	82	90	84

TABLE 3: Matching examples

Keyword	Matched Landxml entities	Score	Correct?
drainage system	Outlet	1.00	yes
	OutletStruct	0.46	
Vertical alignment	roadTerrain	0.58	no
	pointGeometry	0.57	
pavement type	pavementSurfaceType	0.19	yes
	stateType	0.18	
	projectType	0.15	
shoulder	sidewalk	1.00	
	road shoulder	0.61	yes
roadway type	classification	0.56	yes
	RoadSign type	0.44	
aadt	ADT	1.00	yes

buildingSMART (2015). "Ifc overview summary, <http://www.buildingsmart-tech.org/>. Accessed: 2015-10-11.

Cambria, E. and White, B. (2014). "Jumping nlp curves: a review of natural language processing research [review article]." *Computational Intelligence Magazine, IEEE*, 9(2), 48–57.

Costa-Jussa, M. R., FarrÃžs, M., MariÃśo, J. B., and Fonollosa, J. A. (2012). "Study and comparison of rule-based and statistical catalan-spanish machine translation systems." *Computing and Informatics*, 31(2), 245–270.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). "Gate: an architecture for development of robust hlt applications." *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 168–175.

East, E. W. (2007). "Construction operations building information exchange (cobie)." *Report no.*, DTIC Document.

East, E. W., Nisbet, N., and Liebich, T. (2012). "Facility management handover model view."

307 *Journal of computing in civil engineering*, 27(1), 61–67.

308 Eastman, C. (2012). "The future of ifc: Rationale and design of a sem ifc layer. Presentaion
309 at the IDDS workshop.

310 Frantzi, K., Ananiadou, S., and Mima, H. (2000). "Automatic recognition of multi-word
311 terms: the c-value/nc-value method." *International Journal on Digital Libraries*, 3(2),
312 115.

313 Froese, T. (2003). "Future directions for ifc-based interoperability." *ITcon*, 8, 231–246.

314 Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2013). "Semantic measures for the
315 comparison of units of language, concepts or instances from text and knowledge base
316 analysis." *arXiv preprint arXiv:1310.1285*.

317 Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2015). "Semantic similarity from natu-
318 ral language and ontology analysis." *Synthesis Lectures on Human Language Technologies*,
319 8(1), 1–254.

320 Hu, H. (2014). "Development of interoperable data protocol for integrated bridge project
321 delivery." Ph.d., Ph.d. Copyright - Copyright ProQuest, UMI Dissertations Publishing
322 2014 Last updated - 2015-03-18 First page - n/a.

323 inframodel.fi (2014). "Inframodel, <http://www.inframodel.fi/en/>. Accessed: 2015-10-11.

324 landxml.org (2015). "About landxml.org, <http://www.landxml.org/About.aspx>. Ac-
325 cessed: 2015-10-11.

326 Lesk, M. (1986). "Automatic sense disambiguation using machine readable dictionaries: how
327 to tell a pine cone from an ice cream cone." *Proceedings of the 5th annual international*
328 *conference on Systems documentation*, ACM, 24–26.

329 Lima, C., El-Diraby, T., and Stephens, J. (2005). "Ontology-based optimization of knowledge
330 management in e-construction." *Journal of IT in Construction*, 10, 305–327.

331 Lin, J., Hu, Z., Zhang, J., and Yu, F. (2015). "A naturalâĂŘlanguageâĂŘbased approach to
332 intelligent data retrieval and representation for cloud bim." *ComputerâĂŘAided Civil and*
333 *Infrastructure Engineering*.

Marcus, M. (1995). "New trends in natural language processing: statistical natural language processing." *Proceedings of the National Academy of Sciences*, 92(22), 10052–10059.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). "Wordnet: a lexical database for english." *Communications of the ACM*, 38(11), 39–41.

National Institue of Building Sciences (NIBS) (2015). "National bim standard âĂŞ united states version 3." *Report no.*

Navigli, R. (2009). "Word sense disambiguation: A survey." *ACM Computing Surveys (CSUR)*, 41(2), 10.

Nisbet, N. (2012). "Cobie uk: Required information for facility operation." *Report no.*, AEC3 UK Ltd.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). "Feature-rich part-of-speech tagging with a cyclic dependency network." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, 173–180.

U.S. General Services Administration (GSA) (2011). "Gsa building information modeling guide series: 08 âĂŞ gsa bim guide for facility management." *Report no.*

Venugopal, M., Eastman, C. M., Sacks, R., and Teizer, J. (2012). "Semantics of model views for information exchanges using the industry foundation class schema." *Advanced Engineering Informatics*, 26(2), 411–428.

Webster, J. J. and Kit, C. (1992). "Tokenization as the initial phase in nlp." *Proceedings of the 14th conference on Computational linguistics-Volume 4*, Association for Computational Linguistics, 1106–1110.

Yarowsky, D. (1995). "Unsupervised word sense disambiguation rivaling supervised methods." *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 189–196.

Zhang, J. and El-Gohary, N. (2015). "A semantic similarity-based method for semi-automated ifc exension.

Zhao, H. and Kit, C. (2011). "Integrating unsupervised and supervised word segmentation: The role of goodness measures." *Information Sciences*, 181(1), 163–183.