InfraLex: An automatically generated lexicon to support natural language based partial model retrieval for infrastructure projects

Tuyen Le [1], H. David Jeong [2]

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

ABSTRACT

Life cycle project data has been largely available in digital formats to decision makers in the civil sector. Since digital datasets are presented only in computer-readable formats and mostly complicated; data extraction, especially from multiple sources becomes a big burden on the end user. Natural language based data retrieval which allows users to present their data needs in plain English would remove the burden from the user and enable full reuse of life-cycle digital project data. One of the critical requirements for computer to perform this task is digital dictionaries in which meanings of concepts are presented in machine-readable format. This research employs advanced techniques of Natural Language Processing to automatically collect and organize technical terms commonly used in the civil infrastructure sector into a lexical network of terms connecting each other through semantic relations such as synonymy, hypernymy/hypernymy and attribute. Natural Language Processing (NLP) techniques and C-value method are used to automatically detect technical terms from a highway text corpus collected from roadway design guidelines across the U.S. A machine learning model named Skip-gram model is employed to learn the semantic similarity/relatedness between technical terms using the unlabeled corpus as the input data. This model is utilized in a term classification algorithm which can structure related terms into separate groups

[1]Ph.D. Student, Department of Civil, Construction and Environmental Engineering, Iowa State University. Ames, IA 50011, United States. E-mail: ttle@iastate.edu.

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

according to their semantic relations. The developed lexicon has been experimented on the ability of recognizing most semantically similar terms and achieved a precision of 80 percent.
Keywords: Civil infrastructure project, Lexicon, Data retrieval, Natural language interface, NLP, Vector space model

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 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. Thus, there have been apparent demands for an automatic data extraction means that would eliminate the human-relied partial model extracting process.

To address the above demand, a considerable amount of research efforts has been undertaken in both the building and transportation sectors. 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 (Technical Research Center of Finland 2016). 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).

A natural language interface that allows for human-computer interaction in natural language would enable digital data retrieval to overcome the bottle-neck of MVDs and remove the current burden on the end user. One fundamental requirement for such a system is the ability to understand technical terms/keywords since they are the basic unit of natural language and users prefer to use them for obtaining data (Shekarpour et al. 2011). The major obstacle to fulfill the above requirement is the ambiguity issue of technical terms. A technical term in a domain specific document implicitly refers to something that only experts in that field can correctly understand. For example, the term 'roadway type', in general context, can mean the classification of roadways in terms of either material, function or location; but in the highway context, it refers to roadway functional classification. Another issue related to term ambiguity is that two different terms may be used to represent the same concept. For instance, the concept of longitudinal centerline of a roadway has a variety of terms including 'profile', 'crest', 'grade-line' and 'vertical alignment'. Different DOTs may have their own vocabulary system that usually attached as glossary in their documents. In

this case, computer is unable to to exact to map term between data sender and data receiver during the data exchange process if the algorithm is only relying on the data label/name. Addressing those issues will provide a foundation for natural language interfaces to fast and exactly extract data from the complicated sets of data with minimized human intervention and costs.

Recent advances in computer science with considerable improvements have enabled computer to understand human-readable format. This is thanks to the achievement in semantic measure related research which provide infrastructure for computer to present technical keywords in numeric format which can be understood by computer. A large number of methods have been proposed ranging from statistical method to machine learning such as. Distributional model is one of the most common and has been widely used. These achieve high accuracy. The availability of these offer potentials tools for the construction industry to enhance the manual work matching technical keywords in a specific domain to the open data schema. However, like other domains, the highway industry stores knowledge in text documents which are readable to only human. This task aims to transform highway domain knowledge in natural language into a machine-readable format.

This research aims to propose a novel model that can be used to measure the semantic similarity between technical terms in the civil infrastructure domain. In order to achieve that goal, Natural Language Processing (NLP) techniques and C-Value method (Frantzi et al. 2000) are employed to process domain-specific guidelines and extract technical terms commonly used in the civil sector. A matching algorithm implementing the result from the previous step is developed to automatically look for the most nearest entities and attributes in the Landxml schema for a certain keyword. The proposed semantic similarity model and the data mapping algorithm are evaluated by comparing the automatic retrieved data with the manual results from a human for performance assessment. The framework was complied into a Java package and full lexicon datasets which is available at https://github.com/tuyenbk/mvdgenerator.

RELATED RESEARCH

This research employed a hybrid approach which combines a series of techniques related to text analysis and semantic similarity measurement to semantically match user's input keywords to the data entities in the sources schema. Each technique is meant to support each phase of the proposed methodology. The details of research methodology will be presented in the section 3 below. This section presents a the state-of-the-art regarding partial model extraction in the construction industry and a brief introduction to the techniques deployed in the research framework.

Partial model extraction

Methods for extracting partial models for specific use cases can be classified into the following groups ordered by the degree of ease of use for end users: (1) developing a query language specifically for Building Information Modeling (BIM) models, (2) ontology-based query approaches, and (3) user-oriented query methods. The first group aims to tailor the conventional query languages (e.g., SQL, Object Orientation) for extracting information from BIM models. The major focus is on developing spatial filter strategies. Examples of these efforts include the Spatial query language (Borrmann and Rank 2009), QL4BIM Spatio-semantic query language (Daum et al. 2014); graph-based BIM retrieval (Langenhan et al. 2013), and topological querying (Khalili and Chua 2013). The second group is to enhance the human-readability of data schema by utilizing an ontology approach to transform relations among data entities from implicit to explicit. With these semantic presentations, it is easier for end users to read and comprehend a complicated data schema. An extensive number of studies based on this approach have been carried out for various use cases including ontology-driven construction information retrieval for tunnel projects (Min and Zhewen 14), ontology partial BIM model extraction for building projects (Zhang and Issa 2012), ontology-based extraction of construction information (Nepal et al. 2012) and ontology based querying over linked life cycle data spaces (Le and Jeong 2016). The last class of partial model query approaches moves a step further in terms of enhancing the ease of data extraction by

providing query tools that require less effort from users. For example, Won et al. (2013) (Won et al. 2013) proposed a no-schema algorithm that allows for the extraction of IFC instances without using IFC schema or MVD. In addition, a visual BIM query (Wülfing et al. 2014) was also established to visualize query codes. Although significant research efforts have been conducted, there is still a lack of natural language interface platforms that can enable computers to understand and interpret the end user's data interests in the civil infrastructure domain.

Semantic data label matching

In current practices, data input to support a certain data analytics process usually come from multiple resources. These data are stored in different formats and are based upon different vocabularies systems. These inconsistency restricts the ability of data integration and likely leads to semantic ambiguities. In the ontology based data integration and exchange mechanism, ontology serves as a domain data schema. To allow for data exchange or integration, target and source ontology are required to be matched to each other. Matching is the process of find corresponding relationships (e.g. sameAS, isA, etc.) among semantic entities (concept, words, sentences, instances, etc.) between these ontologies (Harispe et al. 2015). These relations are found thanks to the semantic measures which determine the degree of relatedness between concepts (Harispe et al. 2015).

Lacking of an extensive machine-readable dictionary for the civil infrastructure domain

Digital dictionaries, which present definitions of terms in a machine-readable manner, are critical for computer to perform knowledge works such as interpreting users' intention or understanding the meaning behind human-oriented inputs. However, there is still a shortage of such an extensive dictionary for the civil engineering domain. WordNet (Miller 1995) (Miller 1995), which is one of the largest lexicons with over 117,000 synsets, is still generic and not suitable for the highway domain. A few construction domain specific semantic resources have been proposed, for example the Civil Engineering Thesaurus (CET) (Abuzir and Abuzir 2002) (Abuzir and Abuzir 2002), e-Cognos (Wetherill et al. 2002) (Wetherill et al.

2002), and buildingSMART data dictionary (ISO 12006-3) (buildingSMART 2016)(buildingSMART 2016) . Of these knowledge bases, the buildingSMART dictionary is a pioneer semantic database with a long development history of over two decades by the international collaboration of buildingSMART Norway, Construction Specifications Canada (CSC), U.S. Construction Specification Institute (CSI), and STABU Foundation (Hezik 2008) (Hezik 2008). Like other construction specific digital dictionaries, IFD is mainly hand-coded and time consuming; the vocabulary set covers limited number of concepts. Therefore, there is a demand for a computational technique that can automatically develop and maintain these digital dictionaries to keep up with the increasingly arising of new terms.

Lacking of effective semantic mapping algorithms for handling the data ambiguity issue

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) (Zhang and El-Gohary 2015) proposed an algorithm called ZESeM aiming to match a certain keyword to the most semantic nearest IFC entity. The algorithm includes two sequential steps including term-based matching and semantic relation based matching. Since the algorithm accepts matches from the label-based matching step, disambiguation still remains in cases in which the same word form is used for different senses. In addition, ZESeM relies on Wordnet which is a generic lexicon, the applicability would be limited. Lin et al. (2015) (Lin et al. 2015) developed a IFD based framework for BIM information retrieval. IFD Library (International Framework for Dictionaries library), which is developed and maintained by the international buildingSMART, is a dictionary of BIM data terminology that assigns synonyms the same ID. The integration or exchange of data using IDs rather than data names would eliminate semantic mismatch. However, since IFD is a hand-made electronic vocabulary, constructing this e-dictionary is time consuming and therefore it is still very limited to large collection of terms in the construction industry.

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).

Methods for automated measuring semantic similarity

Semantic measurement, which aims to evaluate the similarity or relatedness between semantic units (words, phrases, sentences, concepts, etc.) (Harispe et al. 2015) (Harispe et al. 2015), is one of the main NLP related research topics. The two major approaches for semantic measure include (1) dictionary-based method and (2) distributional method (Harispe et al. 2013) (Harispe et al. 2013). The former method relies on a digital dictionary that consists of terms organized in a lexical hierarchy of semantic relations such as synonym, attribute, hypernym/hyponym, etc. Computational platforms (e.g., information retrieval) built upon such dictionaries are able to fast measure the semantic similarity by computing the distances between words in the hierarchy. Hence, this method would be an ideal solution when digital dictionaries are available. However, digital dictionaries are typically hand-

crafted; they are therefore not available to many domains (Kolb 2008) (Kolb 2008). The latter major method for estimating word similarity is based on the distributional model which represents meanings of words through their contexts (surrounding words) in the corpus (Erk 2012) (Erk 2012). A distributional model stands on the distributional hypothesis that states that two similar terms would occur in the same context (Harris 1954) (Harris 1954). The outcome of this approach is a Vector Space Model (VSM), as illustrated in Figure 1, in which each vector depends on the co-occurrence frequencies between the target word with other words in the vocabulary. The similarity between semantic units in this model is represented by the distance between corresponding points (Erk 2012) (Erk 2012). VSM outperforms the dictionary-based method in terms of time saving as the semantic model can be automatically obtained from text corpus and collecting of these corpus is much easier than manually constructing a digital dictionary (Turney and Pantel 2010) (Turney and Pantel 2010).

The VSM approach has been used in the recent NLP related studies in the construction industry. For example, Yalcinkaya and Singh (2015) (Yalcinkaya and Singh 2015) utilized VSM to extract principle research topics related to BIM from a corpus of nearly 1000 paper abstracts. In addition, this approach was used for information retrieval to search for text documents (Lv and El-Gohary 2015) (Lv and El-Gohary 2015) or CAD documents (Hsu 2013) (Hsu 2013). The increasingly number of successful use cases in the construction industry have evidently demonstrated the promising of the VSM in identifying the semantic similarity between technical terms in order to develop an advanced tools for handling data stored in natural language documents generated through the project life cycle.

Among the methods to develop VSM, Skip-Gram model (Mikolov et al. 2013), which is an un-supervised machine-learning model, outperforms other statistical computational methods in various performance aspects such as accuracy and degree of computational complexity (Mikolov et al. 2013). This machine-learning model learns the semantic similarity between two technical terms through their context similarity. The outcome of the training process is

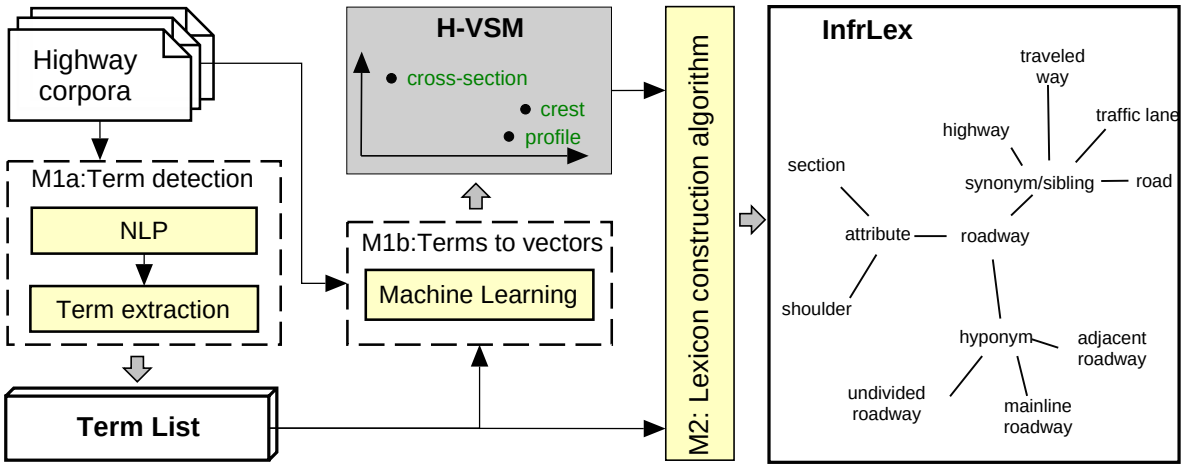

FIG. 1: Overview of the proposed methodology

a set of representation vectors for technical terms.

INFRALEX CONSTRUCTION

Overview of the proposed methodology

The ultimate goal of this research is to build a machine-readable dictionary of technical terms in the infrastructure sector. Conventional methods for building those dictionaries required a huge amount of empirical work, and 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. This research propose a method for automatic construction of a domain thesaurus for the civil infrastructure sector using advanced advanced techniques of Natural Language Processing (NLP).

Figure 1 presents the overview of the methodology for automated generation of a lexicon for infra structure projects (InfraLex). The research framework is consisted of two modules that are: (1) developing a highway term space model (H-VSM), and (2) InfraLex construction. The first module implements several basic NLP techniques (including tokenizing, POS tagging, etc.) and the C-value (Frantzi et al. 2000) method to extract highway related technical concepts from the highway corpora. Skip-gram model, a unsupervised machine learning method proposed by (Mikolov et al. 2013), is then implemented to train the

semantic similarity between technical term using the unlabeled input data of the highway corpora. This training process transforms terms into representation vectors H-VSM. Using this concept vector space, the degree of similarity/relatedness between technical term can be determined; and based on that the list of most semantically similar term for a given term can be obtained. In the second module, a computational algorithm is designed to classify the nearest lists resulted from the H-VSM into lexical groups in accordance to their semantic relations such as synonymy, sibling, hypernymy, hyponymy and attribute. 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.

The process of developing the H-VSM includes the following steps: (1) text document collection, (2) multi-word terms extraction, and (3) semantic similarity training. The sub-sections below discuss the detailed procedures for each step.

Data collection

As mentioned earlier, the H-VSM was trained using a machine learning model which uses text corpus as the training data. A highway corpus was built upon the technical documents collected from multiple sources including textbooks, and highway engineering manuals from the Federal Department of Transportation (DOT) and from 22 distinct State DOTs. The focus of the highway corpora in this research is on the following three project phases: (1) design, (2) construction and (3) asset management. Technical terms in a guidance document in the engineering field are organized in various formats such as plain text, tables, and equations. Since tables and equations are not yet supported by the state-of-the-art NLP techniques, they were removed from the text corpora. The result of data collection is a plain text corpora consisting of 16 million words. This dataset is utilized to extract highway related technical terms which are then trained and converted into vectors.

Multi-word terms extraction

A technical term can be a single word (e.g., roadway, lane, etc.) or be composed of multiple words (e.g., right of way, at grade intersection, etc.). The meaning of multi-word

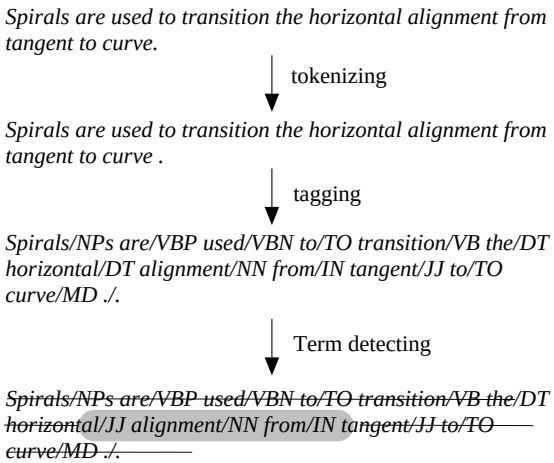

FIG. 2: Linguistic processing procedure to detect technical terms

terms may not be directly interpreted from the meanings of their single words. In order for the Skip-gram model (in the training process) to learn the semantics of multi-word terms, the occurrence of them in the corpus need to be detected and replaced with connected blocks of word members. Figure 2 presents the process of detecting technical terms commonly used in highway technical documents which includes the following steps.

1. **Word tokenizing:** In this step, the text corpus is broken down into individual unit (also called tokens) using OpenNLP Tokenizer.
2. **Part of Speed tagging:** The purpose of this step is to determine the part of speech tag (e.g., noun, adjective, verb, etc.) for each token.
3. **Noun phrase detection:** Linguists argue that a technical term is either a noun (e.g., road) or a noun phrase (NP) (e.g., right of way) that frequently occurs in the domain text documents (Justeson and Katz 1995). Thus, NPs are good multi-word term candidates. Table 1 presents the proposed extraction patterns which are modified from the filters suggested by (1995)(Justeson and Katz 1995) to extract NPs. The first two filters directly detect NPs that occur separately, and the third filter is to count for cases where multiple terms are represented in conjunctions (e.g., 'vertical and horizontal alignment'). For each instance of conjunction, an extra processing

TABLE 1: Term candidate filters

Pattern	Examples
(Adj|N)*N	road, roadway shoulder, vertical alignment
(Adj|N)*N Prep (of/in) (Adj|N)*N	right of way, type of roadway
(Adj|N)* 'and/or' (Adj|N)*N	vertical and horizontal alignment
Note: |, * respectively denotes 'and/or', and 'zero or more'.	

is applied to break it into individual terms. For example, the conjunction 'vertical and horizontal alignment' will become 'vertical alignment' and 'horizontal alignment'. This division process determines the main part ('alignment') which is shared by two terms and the dependent parts ('vertical' and 'horizontal'). This research use the Stanford Dependencies Parsing tool , which is able to analyze dependencies between sentiment units, to split conjunctions phrases into separate phrases.

In addition, in order to avoid the distinguishing between syntactic variants of the same term, for example 'roadway' and 'roadways', term variants will be normalized. The following are three types of syntactic variants and the proposed normalization methods.

- **Type 1** - Plural forms, for example 'roadways' and 'roadway'. The Porter stemming algorithm (Porter 1980), which can assist the automated removal of suffixes is applied on the corpus before extracting NPs.
- **Type 2** - Preposition noun phrases, for example 'roadway type' and 'type of roadway'. In order to normalize this type of variant, the form with preposition needs to be converted into the non-preposition form by removing the preposition and reverse the order of the remaining portions. For example, 'type of roadway' will become 'roadway type'.
- **Type 3** – Abbreviations, such as AADT. A linguistic rule-based method suggested by (Nenadić et al. 2002) will be used to determine the full term for each abbreviation. This method suggested the following abbreviation defini-

tion patterns: (1) left definition pattern – NP (Abbreviation), for example Annual Average Daily Traffic (AADT); and (2) right definition pattern - (Abbreviation) NP, for example (AADT) Annual Average Daily Traffic.

4. **Multi-word term candidate raking and selection:** Multi-word term definition varies between authors, and there a lack of formal rules for defining multi-word term (Frantzi et al. 2000). There are a number of methods for determining termhood (the degree that a linguistic unit is a domain-technical concept) such as TF-IDF (Sparck Jones 1972; Salton and Buckley 1988), C-Value (Frantzi et al. 2000), Termex (Sclano and Velardi 2007). These methods are based on occurrence frequencies of NPs in the corpus. Among these methods, Termex outperformed other methods on the Wikipedia corpus, and C-Value was the best on the GENIA medical corpus (Zhang et al. 2008). This result indicates that C-value method is more suitable for term extraction from a domain corpus rather than a generic corpus. For this reason, the C-value has been widely used to extract domain terms in the biomedical field (Ananiadou et al. 2000), (Lossio-Ventura et al. 2013), and (Nenadić et al. 2002)). Since the corpus used in this research will be mainly collected from technical domain documents, thus C-value would be the most suitable for termhood determination. The C-value measure, as formulated in Equation 1, suggests that the longer a noun phrase is, the more likely that is a term; and the more frequently it appears in the domain corpus, the more likely it will be a domain term.

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

Where:

a is a candidate noun phrase

f is the frequency of a in the corpus

TABLE 2: Examples of extracted terms and evaluation

Term	Termhood	real term?
sight distance	9435.314	yes
design speed	9052.556	yes
additional information	1829.0	no
typical section	1801.0	yes
basis of payment	1762.478	no

Ta is the set of extracted noun phrases that contains a

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

The final ranked list of terms obtained which have C-value greater than 0 was manually evaluated to remove non-terms. Table 2 shows the evaluation result for 5 of the extracted terms. The longer the list is, the more effort required for the evaluation process. Since the the term extraction is based on frequency, the size of term list will be affected if a threshold of frequency is used. With the threshold of 2, the list consists of 112024 terms, and the the list size drops to 8922 when a threshold of 50 is used. Manually reviewing evaluate such a along list is still a challenging task. To minimize human force, the list was evaluated at several ranges of C-values. Precision, which represent percentage of real terms in each group were determined. The results are presented in Figure 3. As show in the figure, precision are relative low for groups with c-values less than 70. To balance between human cost and precision, this research proposes to manually review all the automatically extracted term below the c-value threshold of 70.

Training dataset preparation

The highway text copora collected serves as the source of training dataset for developing the semantic similarity model. The Skip-Gram model requires a set of training data in which the input data is a linguistic unit (word or term) and the output data is a set of context words. In order to collect this training dataset, the unannotated text corpora will be scanned to collect instances of terms and their corresponding context words. Each occurrence of a

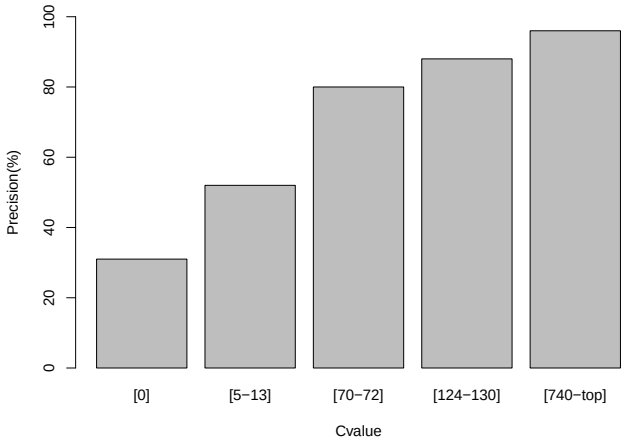

FIG. 3: Precision of term extraction

technical term will correspondingly generate a data point in the training dataset.

Before collecting the training dataset, an additional step is needed to handle the issue related to multi-word terms. Since document scanning is on a word-by-word basis, the corpus must be adjusted so that multi-word terms can be treated like single words. To allow the scanner to collect both full and nested terms, any occurrence of a multi-word term in the corpus is replaced with a single unit that is compiled by connecting all individual words. For instance, 'vertical alignment' will become 'vertical_alignment'.

The adjusted highway corpus is scanned through to find the context words for every occurrence of the technical terms in the vocabulary. The number of context words to be collected is dependent on the window size that limits how many words to the left and to the right of the target word. In the example sentence below, the context of the term 'roadway' with the context window size 10 will be the following word set bike, lane, width, on, a, width, no, curb, gutter. Any word in the stop list (a list of frequent words in English such as 'a', 'an', 'the' that have little meaning) will be neglected in the context set. If the target word is a multi-word term, the set of context words will not include its member words.

"The minimum [bike lane width on a roadway with no curb and gutter] is 4 feet."

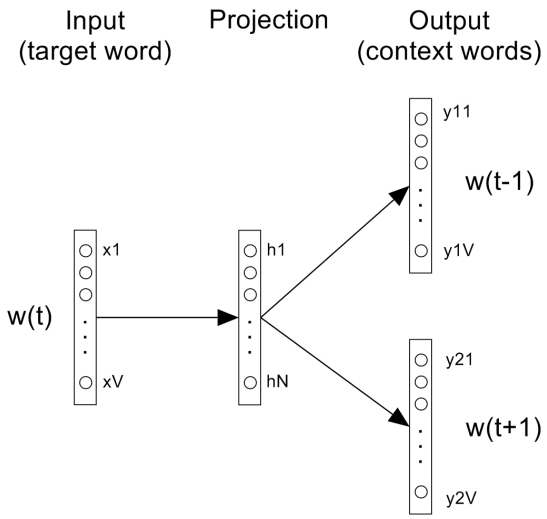

FIG. 4: Skip-gram model

Semantic similarity training

The semantic similarity will be trained using the word2vec module in the Apache Spark MLlib package (Apache.org 2016), an emerging open-source engine, developed based on the Skip-gram neural network model (Mikolov et al. 2013) (Mikolov et al. 2013). The intended parameters used for the training model are presented in Table 3. These values are mainly based on suggestions from the literature. To eliminate data points with low frequency of occurrence that are unlikely to be technical terms, word2vec includes the parameter of minimum occurrence frequency. Any vocabulary with the rate lower than the limit will be ignored. Radim Rehurek, a machine learning consultant company, suggests a range of (0-100)* depending on the data set size. Setting this parameter high will enhance the accuracy, but many technical terms will be out of vocabulary. The preliminary study based on the preliminary corpus with only several millions of words shows that with the frequency of 20, there are very few non-technical terms involved in the training dataset. Hence, with the larger dataset to be collected, this parameter can be higher and up to around 50. The second important parameter is layer size which determines the number of nodes in the hidden layer. This parameter highly affects the training accuracy and processing time. A larger layer size is better in terms of accuracy but this will be paid off by the running time.

TABLE 3: Skip-gram model parameters

Parameter	**Value**
Frequency threshold	50-100
Hidden layer size	100-500
Context window size	5,10,15

Since this research aims to develop a model that can be used for other information retrieval research, the accuracy is the first priority. This parameter may range from 10 to hundreds; in this research, it is expected to in be the range of 100-500. The final major parameter is the context window size. Google suggests the size of 10 for the Skip-gram model. These parameters are subject to be changed so that the best model can be achieved. The effects of these parameters to the performance of the synonym detection application are discussed in Section 4. Figure 5 presents the term space model developed from the training process with the parameters are 50, 300 and 10 respectively. In this model, each technical term collected from technical documents is represented as a vector in a high dimensional space; and the distance between them represents the semantic similarity. The preliminary term space presented in this paper consists of more than six thousand technical keywords. Since the vector space is a multi-dimensional space according to the size of hidden layer. In order to illustrate, present space in 2D graph, PCA (Principle component Analysis) was used to reduce the size to 2 dimensions. The similarity between terms can be measured by the angle between two work representation vectors or the distance between two word points. The following shows two measures of word sense similarity. Table 4 shows an ranked list of near terms obtained from the H-VSM model.

$$cosine\_similarity = \frac{A.B}{||A||.||B||} \tag{2}$$

$$dis\_similarity = \sqrt{(xA_1 - xB_1)^2 + (xA_2 - xB_2)^2 + ... + (xA_n - xB_n)^2} \tag{3}$$

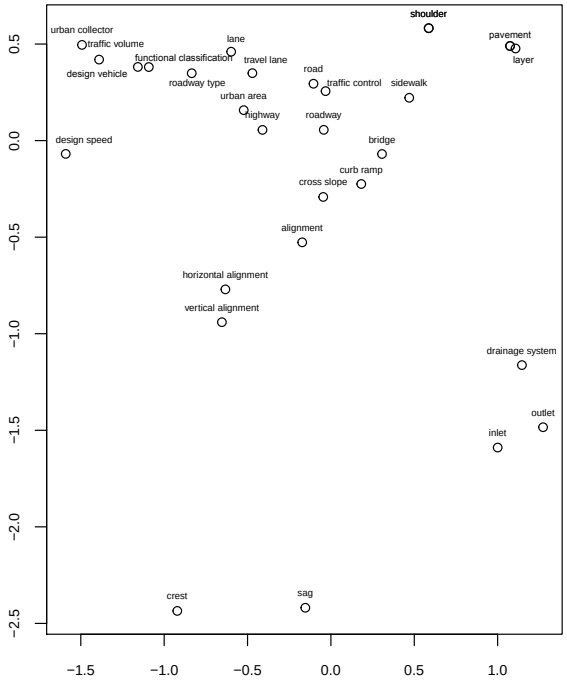

FIG. 5: Highway term space model (H-VSM)

Where: n is the hidden layer size.

Highway lexicon construction

The purpose of this module is to construct a lexicon which is also known as lightweight ontology. A knowledge base typically includes terms and relations. The core relations of an ontology can be classified into the following types: synonym (meaning equivalence), hypernym-hyponym (also known as IS-A or parent-child relation), attribute (concept property), and association (e.g. part-of) (Jiang and Conrath 1997; Lee et al. 2013). If two terms relate each other through these semantic relation would have high similarity. Therefore, the top nearest terms resulted from H-VSM would be a great starting point for detecting relations between technical terms. Table 4 illustrates a list of nearest terms of the term 'roadway'. In this list, true synonyms are highway (1), traveled-way (2) or road(4); roadway-section (3), roadway-shoulder (12) are attributes of the roadway concept; and adjacent-roadway (7), undivided (37) are hyponyms which showing different types of roadway.

The specific objective of this task is to detect relations and based on that rearrange the

TABLE 4: Examples of top nearest terms

Term	Nearests	Cosine	Rank
roadway	highway	0.588	1
	traveled-way	0.583	2
	roadway-section	0.577	3
	road	0.533	4
	traffic-lane	0.524	5
	separating	0.522	6
	adjacent-roadway	0.519	7
	travel-way	0.517	8
	entire-roadway	0.513	9
	...	...	...
	roadway-shoulder	0.505	12
	roadway-cross-section	0.491	18
	undivided	0.452	37
	mainline-roadway	0.450	42

vocabulary constructed in the first module. Algorithm 1 shows the design pseudo code for classify the nearest terms of a given target term. The algorithm utilizes linguistic rules and clustering analysis to reorganize the nearest list into the following three groups including (1) attribute, (2) hyponym and (3) synonym/sibling and functional relation. The algorithm firstly detects terms for the first two categories using linguistic rules. The filter rules to detect these relations are presented in Table 3. For example, with pattern 1, we can infer that Noun1 is a good attribute candidate of concept Noun2; and Noun2 is an attribute of Noun1 in the pattern 2. The remained words will fall into the third group. However, some of them have far or even no relation with the target word. In order to address this issue, this research employed the K-mean clustering algorithm (MacQueen 1967) to split the remained list into three distinct layers based on their similarity score. The terms in the last group are unlikely to be a synonym or sibling and is removed from the nearest list. The outcome of the proposed algorithm is a list of classified nearest term. Table 6 shows one example for the output retrieved from the algorithm.

Algorithm 1 Near term classification algorithm

```
1: Inputs: term t, list of nearest terms N, full list of terms F
2: Output:: Classified list of terms C
3: procedure TERM CLASSIFICATION PROCEDURE
4:     Att ← list of attributes
5:     Hyp ← list of hyponyms
6:     Syn ← list of synonyms
7:     w ← null
8:     for all n ∈ N do
9:         if n contains t then
10:            w ← n
11:        else
12:            for all f ∈ F do
13:                if f contains both n and t then
14:                    w ← f
15:                    Break for
16:    if w matches Attribute pattern then
17:        add w to Att
18:    else if w matches Hyponym pattern then
19:        add w to Hyp
20:    else
21:        add w to Syn
22:    Cluster Syn and discard low relevant terms
```

TABLE 5: Patters to extract attributes and hyponyms

Relation Pattern	**Example**	
Attribute	Noun of Target	the width of the road
	Target Noun	road width, project cost
Hypernym-hyponym	Noun Target	vertical alignment isA alignment

PERFORMANCE EVALUATION

This section presents the performance evaluation of InfraLex in terms of supporting identifying synonyms. In regard of the performance on synonym searching, a gold standard is used. The gold standard is consisting 70 sets of synonyms (both single and multi-word terms) which is examined and extracted from a Wikipedia transportation glossary (Wikipedia 2016). This glossary provided plain text explanation for each term and their synonyms. The automatically identified synonyms which are the top nearest words in the identified synonym

TABLE 6: Examples of top nearest terms

Term	Relation Group	Nearests	Cosine	Rank
roadway	Synonym	highway	0.588	1
		traveled-way	0.583	2
		road	0.533	4
		traffic-lane	0.524	5
		travel-way	0.517	8
	Attribute	separating	0.522	6
		roadway-section	0.577	3
		roadway-shoulder	0.505	12
		roadway-cross-section	0.491	18
	Hyponym	adjacent-roadway	0.519	7
		entire-roadway	0.513	9
		undivided	0.452	37
		mainline-roadway	0.450	42

lists from the algorithm were compared with the true synonyms in the gold standard dataset. The results are measured using the following three measures including recall, precision and f-measure.

$$Recall = \frac{\text{number of correctly matched concepts}}{\text{total concepts}} \tag{4}$$

$$Precision = \frac{\text{number of correctly matched concepts}}{\text{total matched concepts}} \tag{5}$$

$$F - measure = \frac{2.Precision.Recall}{Precision + Recall} \tag{6}$$

Table 7 shows the performance with different training model settings. The baseline model parameter are set 50, 100 and 5 respectively for frequency threshold, hidden layer size and window size. We changed these parameters one by one and remained the other ones to evaluate their effects to the model performance. As presented in the table, the increase of window size to 10 or 15 resulted in the best model which have precision of 81% and

an F-measure of 65%. The change of other parameter does not improve the performance. Especially, the increase of frequency threshold have negative impact.

Table 8 shows the comparison of performance between InfraLex (with 50-100-10 settings) and Wordnet. As shown, InfraLex outperformed Wordnet in all aspect, and the combined F-measure is significantly improved (65 compared to 52). The biggest contribution to the improvement is the recall value which represents the coverage of domain vocabulary.

Future research is till needed to enhance the performance of InfraLex, especially the recall measure. One potential reason for the low recall is due to the the training data size. This model currently is based on the data training set consisting of only 16 million words. In order to enhance the accuracy, the data training set needs to be extended in size and in the numbers of disciplines related to highway projects.

DISCUSSIONS

the current study as a number of limitations. first, the highway corpus is till relative small with only 160 million words, compared to corpus in other domain, this is relative small. since the the recall largely depends on the corpus size. the larger corpus the more vocabulary can

TABLE 7: Effects of training parameters on performance of synonym matching

Parameter	**Model**	**Precision (%)**	**Recall(%)**	**F (%)**
Baseline	50-100-5	79	53	63
Window size	**50-100-10**	**81**	**54**	**65**
	50-100-15	81	54	65
Frequency threshold	75-100-5	74	50	60
	100-100-5	77	51	62
Hidden layer size	50-200-5	79	53	63

TABLE 8: Comparison of synonym matching performance between Wordnet and InfraLex

Lexicon	**Precision (%)**	**Recall(%)**	**F (%)**
Wordnet	76	40	52
InfraLex	**81**	**54**	**65**

be extracted and larger size means some more technical terms can be covered.

This method is broad for nlp and linguistics tools 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. Emerging of text mining and natural language processing related research. the following application can be used: (1) text mining to extract information from text of which the basic linguistic unit is term. understanding of term meaning through it's synonyms, hyponyms and attributes will allow computer to extraction information precisely. (2) natural language interface. the current interaction between computer and user, for example in data extraction require big effort from user, a natural language interface which can understand natural language user's input will allow the automatic extraction of information. data information retrieval query by query their sysnonym, information classification. use for disambiguation task to interpret user's intention.

CONCLUSIONS

Data manipulation and retrieval from multiple sources is challenging task due to inconsistency of data format and terminology. The contribution of this study This research develops a digital dictionary of highway related technical terms which can enable computer understand semantic meanings of terms. this research employs advanced NLP techniques to extract technical terms from a highway text corpus cosisting of 160 million words built on design manuals from 22 State DOT across the U.S. Machine learning were used to train semantic similarity between technical terms which can be used to retrieve a list of most semantically similar terms. A algirthm was designed to 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 comparing the resulted obtained from the InfraLex model and a man-crafted gold standard. The result shows the accuracy of over 80 percent. the best model achiveed with the window size of 10. The accuracy is low due to the size of the training data. Future research will be conducted to expan the highwah corpus to further disciplines such as asset management, trasportation operation.

the research open new gate for computerational tools regarding natural lanauge processing in the highway sectors. the infraLex would enable computer to understand terms and consequently transform the way human interact with computer. once the ease of use when natural language can be used, the seamless reuse of lifecyle data would be achieved in the highway sector.

REFERENCES

Abuzir, Y. and Abuzir, M. O. (2002). "Constructing the civil engineering thesaurus (cet) using the theswb." *Computing in Civil Engineering*.

Ananiadou, S., Albert, S., and Schuhmann, D. (2000). "Evaluation of automatic term recognition of nuclear receptors from medline." *Genome Informatics*, 11, 450–451.

Apache.org (2016). "Machine learning library (mllib), <https://spark.apache.org/docs/1.1.0/mllib-guide.html> (March).

Borrmann, A. and Rank, E. (2009). "Query support for bims using semantic and spatial conditions." *Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies: Concepts and Technologies*, 405.

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

buildingSMART (2016). "Data dictionary, <http://www.buildingsmart.org/standards/standards-library-tools-services/data-dictionary/>. Accessed: March 15, 2016.

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., Farrús, M., Mariño, 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.

Daum, S., Borrmann, A., Langenhan, C., and Petzold, F. (2014). "Automated generation of building fingerprints using a spatio-semantic query language for building information models." *eWork and eBusiness in Architecture, Engineering and Construction: ECPPM 2014*, 87.

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

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

Erk, K. (2012). "Vector space models of word meaning and phrase meaning: A survey." *Language and Linguistics Compass*, 6(10), 635–653.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). "Automatic recognition of multi-word terms: the c-value/nc-value method." *International Journal on Digital Libraries*, 3(2), 115–130.

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

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2013). "Semantic measures for the comparison of units of language, concepts or instances from text and knowledge base

analysis." *arXiv preprint arXiv:1310.1285*.

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

Harris, Z. S. (1954). "Distributional structure." *Word*.

Hezik, M. (2008). "Ifd library background and history." *The IFD Library/IDM/IFC/MVD Workshop*.

Hsu, J.-y. (2013). "Content-based text mining technique for retrieval of cad documents." *Automation in Construction*, 31, 65–74.

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

Jiang, J. J. and Conrath, D. W. (1997). "Semantic similarity based on corpus statistics and lexical taxonomy." *arXiv preprint cmp-lg/9709008*.

Justeson, J. S. and Katz, S. M. (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text." *Natural Language Engineering*, 1(01), 9–27.

Khalili, A. and Chua, D. (2013). "Ifc-based graph data model for topological queries on building elements." *Journal of Computing in Civil Engineering*, 0(0), 04014046.

Kolb, P. (2008). "Disco: A multilingual database of distributionally similar words." *Proceedings of KONVENS-2008, Berlin*.

landxml.org (2015). "About landxml.org, <http://www.landxml.org/About.aspx>. Accessed: 2015-10-11.

Langenhan, C., Weber, M., Liwicki, M., Petzold, F., and Dengel, A. (2013). "Graph-based retrieval of building information models for supporting the early design stages." *Advanced Engineering Informatics*, 27(4), 413–426.

Le, T. and Jeong, H. D. (2016). "Interlinking life-cycle data spaces to support decision making in highway asset management." *Automation in Construction*, 64, 54–64.

572 Lee, T., Wang, Z., Wang, H., and Hwang, S.-w. (2013). "Attribute extraction and scor-
573 ing: A probabilistic approach." *Data Engineering (ICDE), 2013 IEEE 29th International*
574 *Conference on*, IEEE, 194–205.

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

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

580 Lin, J.-R., Hu, Z.-Z., Zhang, J.-P., and Yu, F.-Q. (2015). "A natural-language-based approach
581 to intelligent data retrieval and representation for cloud bim." *Computer-Aided Civil and*
582 *Infrastructure Engineering*.

583 Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2013). "Com-
584 bining c-value and keyword extraction methods for biomedical terms extraction."
585 *LBM'2013: 5th International Symposium on Languages in Biology and Medicine*,
586 http://lbm2013.biopathway.org/. Computer Science [cs]/Bioinformatics [q-bio.QM] Life
587 Sciences [q-bio]/Quantitative Methods [q-bio.QM] Computer Science [cs]/Document and
588 Text ProcessingConference papers.

589 Lv, X. and El-Gohary, N. M. (2015). "Semantic annotation for context-aware information
590 retrieval for supporting the environmental review of transportation projects.

591 MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observa-
592 tions." *Proceedings of the fifth Berkeley symposium on mathematical statistics and proba-*
593 *bility*, Vol. 1, Oakland, CA, USA., 281–297.

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

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

598 Miller, G. A. (1995). "Wordnet: a lexical database for english." *Communications of the ACM*,

599 38(11), 39–41.

600 Min, H. and Zhewen, H. (14). "Ontology-driven tunnel construction information retrieval and
601 extraction." *Control and Decision Conference (2014 CCDC), The 26th Chinese*, IEEE,
602 4741–4746.

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

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

607 Nenadić, G., Spasić, I., and Ananiadou, S. (2002). "Automatic acronym acquisition and term
608 variation management within domain-specific texts." *Third International Conference on*
609 *Language Resources and Evaluation (LREC2002)*, 2155–2162.

610 Nepal, M. P., Staub-French, S., Pottinger, R., and Zhang, J. (2012). "Ontology-based feature
611 modeling for construction information extraction from a building information model."
612 *Journal of Computing in Civil Engineering*, 27(5), 555–569.

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

615 Porter, M. F. (1980). "An algorithm for suffix stripping." *Program*, 14(3), 130–137.

616 Salton, G. and Buckley, C. (1988). "Term-weighting approaches in automatic text retrieval."
617 *Information processing & management*, 24(5), 513–523.

618 Sclano, F. and Velardi, P. (2007). *Termextractor: a web application to learn the shared*
619 *terminology of emergent web communities*. Springer, 287–290.

620 Shekarpour, S., Auer, S., Ngomo, A.-C. N., Gerber, D., Hellmann, S., and Stadler, C. (2011).
621 "Keyword-driven sparql query generation leveraging background knowledge." *Web Intelli-*
622 *gence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International*
623 *Conference on*, Vol. 1, IEEE, 203–210.

624 Sparck Jones, K. (1972). "A statistical interpretation of term specificity and its application
625 in retrieval." *Journal of documentation*, 28(1), 11–21.

626 Technical Research Center of Finland (2016). "Finnish inframodel application documentation
627 for landxml v1.2 - version 4:2016, <http://cic.vtt.fi/inframodel/>. Accessed: April 09,
628 2016.
629 Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). "Feature-rich part-of-
630 speech tagging with a cyclic dependency network." *Proceedings of the 2003 Conference of*
631 *the North American Chapter of the Association for Computational Linguistics on Human*
632 *Language Technology-Volume 1*, Association for Computational Linguistics, 173–180.
633 Turney, P. D. and Pantel, P. (2010). "From frequency to meaning: Vector space models of
634 semantics." *Journal of artificial intelligence research*, 37(1), 141–188.
635 U.S. General Services Administration (GSA) (2011). "Gsa building information modeling
636 guide series: 08 – gsa bim guide for facility management." *Report no.*
637 Venugopal, M., Eastman, C. M., Sacks, R., and Teizer, J. (2012). "Semantics of model
638 views for information exchanges using the industry foundation class schema." *Advanced*
639 *Engineering Informatics*, 26(2), 411–428.
640 Webster, J. J. and Kit, C. (1992). "Tokenization as the initial phase in nlp." *Proceedings of*
641 *the 14th conference on Computational linguistics-Volume 4*, Association for Computational
642 Linguistics, 1106–1110.
643 Wetherill, M., Rezgui, Y., Lima, C., and Zarli, A. (2002). "Knowledge management for the
644 construction industry: the e-cognos project.
Wikipedia (2016). "Glossary of road transportation terms,
<https://en.wikipedia.org/wiki/Glossary$_o f_r oad_t ransport_t erms$>.*Accessed* :
*April*11, 2016.
645 Wülfing, A., Windisch, R., and Scherer, R. (2014). "A visual bim query language." *eWork and*
646 *eBusiness in Architecture, Engineering and Construction: ECPPM 2014*, 157.
647 Won, J., Lee, G., and Cho, C. (2013). "No-schema algorithm for extracting a partial model
648 from an ifc instance model." *Journal of Computing in Civil Engineering*, 27(6), 585–592.
649 Yalcinkaya, M. and Singh, V. (2015). "Patterns and trends in building information modeling

(bim) research: A latent semantic analysis." *Automation in Construction*, 59, 68–80.

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.

Zhang, L. and Issa, R. R. (2012). "Ontology-based partial building information model extraction." *Journal of Computing in Civil Engineering*, 27(6), 576–584.

Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). "A comparative evaluation of term recognition algorithms." *LREC*.

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