NLP-based approach to structuring heterogeneous terms for unambiguous exchange of highway data

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

ABSTRACT

The inconsistency of data terminology due to the fragmented nature of the civil infrastructure industry has imposed big challenges on integrating digital data from distinct sources to support decision making in asset management. The issue of data ambiguity may lead to a lack of common understanding to the same data between the sender and receiver. While the heterogeneity of data formats has been well addressed thanks to the availability of various international neutral data standards such as LandXML and TransXML; the heterogeneity of terms where various terms used from the same concept still has been neglected by the domain researchers. This paper presents a novel methodology to construct an automatically-generated lexicon, namely InfraLex, that formally organizes civil infrastructure technical terms in a lexical hierarchy manner. The lexicon severs as a digital dictionary of domain terms which would enable data integration systems to understand the meaning of a data representation, and helps avoid the mismatch of data. Natural Language Processing (NLP) techniques and the C-value method are used to detect technical terms from a highway text corpus collected from roadway design guidelines across the State Departments of Transportation. A model for measuring term similarity is trained using the Skip-gram model which uses the corpus as the training dataset. This semantic model is then utilized by a term classification algorithm that organizes related terms into separate groups according to their semantic relations. The developed lexicon was evaluated by conducting an experiment

[1]Ph.D. Candidate, 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.

comparing the automatically-identified synonyms with a human-constructed synonym set. The result shows that the proposed model achieved a precision of over 80 percent.

Keywords: Civil infrastructure project, Lexicon, Data sharing, Semantic Interoperability, NLP, Vector Space Model

INTRODUCTION

The implementation of advanced computerized technologies such as 3D modeling and Geographic Information System (GIS) throughout the life cycle of a civil infrastructure project has allowed a large portion of project data to be available in digital format. The efficiency improvement in sharing these data between project participants and stages, will in turn, translate into increased productivity, efficiency in project delivery and accountability. However, a highway asset as a whole has not yet fully benefited from the potentials of digital datasets as an accessible, reusable and reliable information source for life-cycle decision making due to the interoperability issue. A study by the National Institute of Standard and Technology (NIST) reported an estimated inadequate interoperability cost in the U.S. capital facility industry at over $15.8 billion per year (Gallaher et al. 2004). This study also pointed out that two-third of those cost occurs during the operation and maintenance stages, and the largest cost item is the labor work for finding, verifying, and transferring facility and project information into a useful format. This finding indicates that the lack of readiness for downstream phases to directly use the transferred digital project data generated from upstream design and construction stages results in high operational costs. Since the roadway sector, which is one of the largest domains in the construction industry, have not yet successfully facilitated a high degree of interoperability (Lefler 2014); a huge cost saving would be achieved if roadway data is seamlessly shared through the project life cycle and among state and local agencies.

Semantic interoperability, which relates to the issue whereby two computer systems may not have the same understanding to the same piece of data, is a radical barrier to computer-to-computer data exchange. Due to the fragmented nature of the infrastructure domain,

data representation/terminology varies between phases, stakeholders, or geographic regions (counties, states, etc.). Retrieving the right pieces of data in such a heterogeneous environment becomes increasingly complex (Karimi et al. 2003). Polysemy and synonymy are two major linguistic obstacles to semantic integration and use of a multitude of data sources (Noy 2004). Polysemy refers to cases when a unique term has several distinct meanings. For example, the term *roadway type* can either mean *material classification* or *functional classification* of roadways. In contrast, synonymy is associated with the heterogeneity of terms used for the same concept. For instance, the longitudinal centerline of a roadway has various terms including 'profile', 'crest', 'grade-line' and 'vertical alignment'. Simply mapping of data names will likely lead to the failure of data extraction, or use of wrong data. Thus, addressing the terminology ambiguity due to the semantic heterogeneity issue becomes crucial to ensure the common understanding on the same dataset between software applications and guarantee the extraction of right data and proper integration of data from multiple sources.

Terminology transparency through domain knowledge resources like glossaries, taxonomies, ontologies and data dictionaries is identified as an driver of semantic interoperability (Ouksel and Sheth 1999). A digital dictionary explains domain concepts in a machine-readable format would allow computer systems to precisely understand meanings of terms. Thus, data mismatch when unifying isolated data sources would be eliminated. A plethora of construction specific semantic resources have been proposed; for example e-Cognos (Wetherill et al. 2002), and buildingSMART data dictionary (ISO 12006-3) (buildingSMART 2016), freight data (Seedah et al. 2015a), highway asset management (El-Diraby and Kashif 2005), infrastructure (Osman and Ei-Diraby 2006). and various data integration and sharing based ontology have been proposed for integration of heterogeneous sources, for instance, ones developed by (Buitelaar et al. 2008) and (Seng and Kong 2009). However, developing an ontologies still a bottleneck to semantic interoperability and ambiguous data sharing and reuse in the infrastructure industry. Although a large number of ontologies available, a large portion

of the civil infrastructure related concepts and synonyms are not yet included investigated. This is due to the reliance on manual approaches to ontology construction which is ah-hoc, laborious and time-consuming. These ontologies are mainly hand-coded/hand-crafted, laborious/tedious, time-consuming and;. These ontologies were developed by manually review domain knowledge and select representation terms for the concepts and assign relations for each pair of terms. Several software applications such as Portege like Computer-Aid-Drawing can assist in hand constructing of ontology using computer, but mostly all of task performed by developers. There is a need for more rigorous computational techniques that can allow for automated extraction of data with minimized human intervention (Mounce et al. 2010), such that digital dictionaries can be quickly constructed to meet a specific domain and to keep up with the continued/sustainable growth of new terms arisen along with new knowledge and technologies.

Recent achievements in accuracy and processing time of advanced Natural Language Processing (NLP) techniques which employ statistics and machine learning have driven text mining and cognitive recognition research to a new era. There is a rich set of NLP tools supporting text processing ranging from Part of Speech (POS) tagging (Toutanova et al. 2003; Cunningham et al. 2002) for single linguistic units, to Dependency parser (Chen and Manning 2014) for relationships among units. These basic NLP techniques have been applied in various computational platforms that can support deep linguistic analysis at the semantic level of terms such as Word2vec (Mikolov et al. 2013), and Glove (Pennington et al. 2014). The availability of these NLP tools offers great potentials for the construction industry where most of the domain knowledge resources are in text documents (e.g., design guidelines, specifications, etc.). The implementation of NLP will allow for a fast translation of domain knowledge into computer-readable format which is required for a machine-to-machine based data exchange.

This paper presents the an an automated approach using NLP to process of translating text-based domain knowledge into an extensive lexicon, namely InfraLex, for the domain

of civil infrastructure. The lexicon formally organizes civil infrastructure technical terms in a lexical hierarchy manner that can serve as the core dictionary in a data integration system. In order to achieve that goal, several Natural Language Processing (NLP) techniques and the C-value method (Frantzi et al. 2000) are used to detect technical terms from a highway text corpus collected from roadway design guidelines across the Sate Departments of Transportation. A model for measuring term similarity is trained using the Skip-gram model (Mikolov et al. 2013) which uses the highway corpus as the training dataset. This semantic model is then utilized by a proposed term classification algorithm which reorganizes related terms into separate groups according to their semantic relationships including synonym, hyponym and functional relation. A Java package and a lexicon dataset result from the study can be found at https://github.com/tuyenbk/mvdgenerator.

The paper is organized as follows. This section presents the background and rationale for the study. Section 2 discusses the underling knowledge supporting the study and the gap of knowledge. Sections 3 and 4 respectively describe the methodology employed to develop InfraLex and the performance evaluation results. Research limitations and potential applications are discussed in Section 5. The final section concludes the research with discussions on the major findings and future research.

BACKGROUND

This section presents the state-of-the-art regarding NLP and methods to measure semantic similarity which is followed by a review of related research and the gap of knowledge associated with data disambiguation in the civil infrastructure sector.

Natural Language Processing

NLP is a research area developing techniques that can be used to analyze and derive value information from natural languages like text and speech. The major applications of NLP include translation, information extraction, opinion mining (Cambria and White 2014), etc. These applications are embodied by a rich set of NLP techniques ranging from syntactic processing at the word individual level such as Tokenization (breaking a sentence into indi-

vidual tokens) (Webster and Kit 1992; Zhao and Kit 2011), Part-of-Speech (POS) tagging (assigning tags like adjective, noun, verb, etc. to each token of a sentence) (Toutanova et al. 2003; Cunningham et al. 2002), and Dependency parser (relationships between linguistic units) (Chen and Manning 2014), to the semantic level like 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. Rule-based methods, which rely solely on hand-coded rules, are not able to fully cover all complicated sets of human grammatical rules (Marcus 1995); and their performance is, therefore, relatively low. In contrast, the ML-based approach is independent of languages and linguistic grammars (Costa-Jussa et al. 2012) as linguistics patterns can be fast learned from even un-annotated training examples. Thanks to its impressive out-performance, NLP research is shifting to statistical ML-based methods (Cambria and White 2014).

Vector Representation of Word Semantics

Measuring of semantic similarity, which is one of the main NLP-related research topics, aims to determine how much two linguistic units (e.g., words, phrases, sentences, concepts) are semantically alike. For example, a *bike* might be more similar to a *car* than to *gasoline*. The state-of-the-art methodology for measuring similarity can be divided into two categories that are (1) thesaurus-based methods and (2) vector space models (VSM) (Harispe et al. 2013). The former method relies on a hand-coded 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). The latter method, on the other hand, analyzes meaning of words or phrases by considering the statistics of occurrence words and their contexts in natural language text documents. VSM outperforms the dictionary-

based method in terms of time saving as a semantic model can be automatically obtained from a text corpus and corpus collecting is much easier than manually constructing a digital dictionary (Turney and Pantel 2010).

VSM estimates semantic similarity based on the *distributional model* which represents the meaning of a word through its context (co-occurring words) in the corpus (Erk 2012). The distributional model stands on the *distributional hypothesis* that states that two similar terms tend to occur in the same context (Harris 1954). The outcome of this approach is a Vector Space Model (VSM), in which each vector represents a word in the vocabulary. The similarity between semantic units in this model is represented by the distance between the corresponding points (Erk 2012). The conventional method to construct vector representation of semantic units is to use the 'word-context' matrix that shows how frequent a word A is a context of word B in a given text corpus (co-occurrence of words with one another in the corpus). the raw frequencies are factorized using a weighting method. each row in the factorized matrix yields a vector representation. Pointwise Mutual Information (PMI) (Church and Hanks 1990) or it's variant, Positive PMI (PPMI) is a popular measure to transform raw frequencies to statistical probabilities which determine whether the co-occurrence result from the natural random or special semantic relations. The more advanced approach is to use machine learning to train vector representation of terms. One of the recent neural language models is Skip-gram (Mikolov et al. 2013), which is an un-supervised machine-learning model that predicts the context words of a given input word, and the objective is to minimize the error between the predicted and the actual context vectors. An alternative un-supervised machine learning is Glove which is developed by (Pennington et al. 2014) with the objective is that the dot product of two vector presentations equals the logarithm of the word's probability of co-occurrence. The major difference between these two models is that Skip-Gram model trained local context data within a context window, the Glove trains on the global co-occurrence statistics. There are contradict claims on the best model when the authors of these two learning model both claimed the outperformance of their models to

the state of the art. An independent study conducted by (Levy et al. 2015) benchmarking with the state-of-the-art models in various tasks and golden standards shows that Skip-gram outperforms Glove in every experiment and is the winner in most of the tasks, especially on WordSim Similarity dataset. The best score of Skip-gram is .793, while the highest score of PPMI is 7.55 and the glove achieve the highest score of .725. The outperformance of the Mikolov model on the similarity task is confirmed in another benchmarking study carried out by (Hill et al. 2015) on the SimLex-999.

The VSM approach has been progressively implemented in the recent NLP related studies in the construction industry. For example, Yalcinkaya and Singh (2015) utilized VSM to extract principle research topics related to BIM from a corpus of nearly 1,000 paper abstracts. In addition, this approach was used for information retrieval to search for text documents (Lv and El-Gohary 2015) or CAD documents (Hsu 2013). The increasingly number of successful use cases in the construction industry has evidently demonstrated that the VSM method can successfully identify the semantic similarity between data labels which is critical to address the issue of semantic interoperability in sharing digital data across the life cycle of an infrastructure project.

A review of semantic interoperability efforts

Digital dictionaries, which present definitions of terms in a machine-readable manner, are critical for a machine to perform knowledge works such as interpreting users' intention or understanding human-oriented inputs. There are various forms of digital dictionaries for instance,glossaries, thesauruses, taxonomy, ontology. A plethora of construction specific semantic resources, ontologies and taxonomies have been proposed. e-COGNOS (Consistent knowledGe management across proOjects and between enterpriSes in the construction domain), project is a pioneer effort aims to formulate construction knowledge in ontology manner (Wetherill et al. 2002; **?**). This is high level description of a construction project with major concepts include technical topics, actors, resources, products, processes, project and relations among these concepts, and relevant concept. The ontology developers reviewed

existing taxonomies such as BS61000, UniClass, IFC, and end users' documents, construction specific standards and interact with the end users to develop an ontology consists of a terms, relations among concepts, taxonomy enriched with equivalent terms. The end industry experts were invited to validate the developed ontology through questionnaires the the coverage of domain knowledge and terms. This ontology development approach also adopted in a numerous number of research which developed a variety of ontologies for different processes and sub-domains during the life cycle of an construction projects across many construction sectors such as a taxonomy for construction project concepts (El-Diraby et al. 2005), freight ontology (Seedah et al. 2015a), highway (El-Diraby and Kashif 2005), infrastructure (Osman and Ei-Diraby 2006).BS6100-4:2008 (bs6 2008). Ontology allow transparency of definition of a concept, but they are not yet able to address the terminology heterogeneity regarding the issue of synonyms and synonymy. the conversions between terms in heterogeneous systems isolated systems which may use equivalent terms.

Another branch of research area that are more focused on the semantic heterogeneity issue rather standardizing knowledge representation like ontologies, data dictionaries aims to develop data dictionaries. buildingSMART data dictionary (bsDD) (ISO 12006-3) (buildingSMART 2016) is a database for conversions between international terms for different language, systems, term variants, synonym. each concept for instance "slab" will be assigned a unique ID, and all of different equivalent terms in english such as slab or in other language like "ventana" in Spanish will be assigned to this ID. 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). IFD (International Framework for Dictionaries library), which is developed and maintained by the international buildingSMART, is a mechanism for integrating or exchange between those BIM models developed using bsDD that may use heterogeneous data terminology. IFD matching using IDs rather than data names would minimize the occur-

rence of semantic mismatches. IFD separate concepts from names or words referring the concept. To enhance the awareness of the semantic heterogeneity in the transportation sector, a research conducted by (Walton et al. 2015; Seedah et al. 2015b) have surveyed on a wide range of freight related databases and and synthesize the difference in data definition among databases. the authors proposed a data element classification framework which classifies freight data into 9 distinct roles/groups including time (e.g., year, month, time, day), place (city name, state, area, population), commodity (liquid, bulk, value, weight), link (roadway name, width, length), mode (truck, rail, unit train, vehicle class), industry (company name, sales, employee number, annual payroll), event (accident, number of fatalities), human (officer, drunk driver, driver age). based on this classification within each database, the related data entities across separate databases can be identified. the authors classified the data element for 28 public and commercial freight data sources with the total element over 6,300 data elements. using this classified system shows the semantic heterogeneity among systems when these data sources using distinct terms for the same concepts. review metadata and provide examples of cases where difference in data element definition among individual sources, over time periods, traffic volume count (it could either be AADT (annual average daily traffic), or AAWDT (average annual weekday traffic)). An automated approach, ThesWB to construct a digital dictionary was the Civil Engineering Thesaurus (CET) from html web pages (Abuzir and Abuzir 2002). In addition, traditional ontology therefore, are not address the issue of linguistic mismatches like polysemy and synonymy. However, there is still a shortage of such an extensive dictionary that can used for the civil engineering domain. Like other construction specific digital dictionaries, buildingSMART dictionary is mainly/empirical approach hand-coded and time consuming; the vocabulary, therefore, is still relatively limited. Therefore, it is required to develop a handy computational technique that can assist in developing and maintaining these digital dictionaries.An automated approach, ThesWB to construct a digital dictionary was the Civil Engineering Thesaurus (CET) from html web pages (Abuzir and Abuzir 2002). this method used only

syntactic patterns to extract hierarchical relations between terms which drawback is reported as weak (Marcus 1995). (Rezgui 2007) semi-auotmated ontology learning from domain text corpus ifc learning from text using co-occurrence based measure of tf-idf to evaluate identify potential relations between the new extracted keywords and concepts in the existing ontology. potential pairs of related concepts are categorized by domain experts. identify important concepts using if-idf and relations among new concepts and existing concepts in the existing ontologies. classification of relations: (1) speciallization/generalization (based on the same name or a common stem, for instance, speration wall and structural wall), (2) composition/aggregation relationship (e.g., door is an aggregation of a frame, a handle, etc.), and (3) semantic relationship between concepts (e.g., a beam support a slab). The last two relations are identified using a semi-automated approach in which potential relations are detected first based on co-occurence in the same sentence using 'Metric Clusters' method and then they are classified by domain expert to assign/define relation names between two potential related concepts. this research not yet capture the synonym relations, future research is needed to capture this relations. (Hsieh et al. 2011) automated construction of ontology from engineering handbooks, concept relations, concept hierarchy without extraction the lexical relations like synonyms, hyponyms. (Niu and Issa 2015) American Institute of Architects (AIA) Document A201-2007 documents to proposed a taxonomy development approach used utilize OAT, a plugin in GATE (General Architecture for Text Engineering) develop taxonomy associated with construction contracts. developer needs to read and understand all of the contract clauses. the developer would use OAT (Ontology Annotation Tool) interface to manually annotate class names/properties/ instance for each concept in the contract clause. it challening the the number of text increase and hard to manage a long list of annotation tags. deveopler must a industry expert who can understand and the contract clauses and undertand meaning of term so that the annotation can be precisely annotated. in addition, extrinsic validation with the involvement of domain experts to validate the developed taxonomy. They, semantic resources, are mainly hand-coded, laborious/tedious, time-consuming

and, and become a bottleneck therefore, still cover only a small portion of the civil infrastructure related concepts and synonyms are not yet included. But the civil infrastructure are complex with various domain areas, manual process is not able to handle all of these. In addition, ontology These ontologies were developed by manually review domain knowledge and select representation terms for the concepts and assign relations for each pair of terms. some software applications such as Portege can assist in constructing ontology in digital, but mostly all of task performed by developers. As discussed above, since the manual process of dictionary development, the domain dictionaries are currently far below the required scope of vocabularies and achieving the desired size is challenging. Thus, an automated approach that can allow for fast development of highway lexicon instead of relying on hand-coded resources is needed. while the construction projects, especially the civil infrastructure projects, there enormous of proceses during the life cycle of the project, with the involement of varous project stake holders and organizations, and disciplines, the current process is ad-hoc, developing an extensive taxonomy for the whole rpoject life cycle for differnt types of assets with consideration the heterogeneity of terms between organizations become challenging. there is demand for a more rigorous method to classify these terms in a formal manner in which both computer and human readable. ontology based on the interact with domain experts of a specific regions the name of concepts is from that regions, thus, equivalent terms are not addressed.

PROPOSED METHODOLOGY TO AUTOMATED CLASSIFICATION OF HIGHWAY TERMS

Overview of the proposed methodology

The ultimate goal of this research is to construct a machine-readable dictionary of American-English technical terms, named InfraLex, for the infrastructure sector. Figure 1 presents an overview of the methodology proposed to develop InfraLex. The research framework consists of two major modules that are to: (1) train a highway vector space model (H-VSM), and (2) develop an algorithm integrating H-VSM and various linguistic patterns

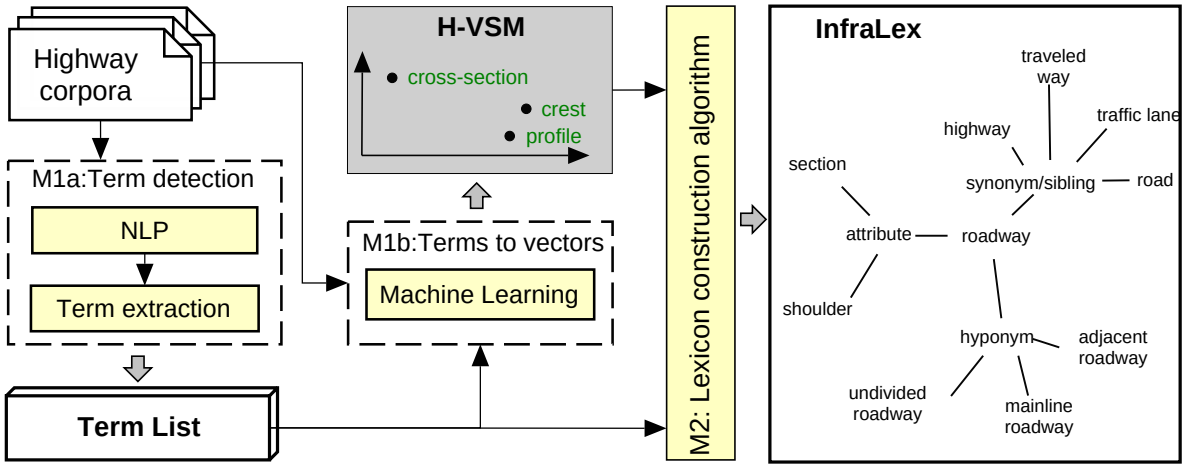

FIG. 1: Overview of the proposed methodology

to construct InfraLex. The first module implements several basic NLP techniques (including tokenizing, POS tagging, etc.) and C-value method (Frantzi et al. 2000) to extract highway related technical terms from a highway corpora. The Skip-gram model, an unsupervised machine learning platform proposed by Mikolov et al. (2013), is then employed to train the semantic similarity between technical terms. The model uses the unlabeled highway corpora as the training dataset. This training process transforms the identified terms into representation vectors in a coordinate space model named H-VSM. Using this term vector space, the similarity degree between technical terms can be determined; and based on that the list of nearest terms for a given term can be obtained. The second module designs an algorithm for identifying the relation (e.g., synonymy, hypernymy, hyponymy, or attribute) between each item in the nearest list and the target term. The InfraLex lexicon is finally constructed by organizing the domain vocabulary into a network of terms which link to each other through the identified semantic relations. Specifically, the procedure followed to compile the InfraLex dictionary is comprised of the following steps: (a) collect highway technical documents to compose a domain corpus; (b) extract the multi-word terms from the highway corpus; (c) prepare the training dataset for training the H-VSM model; (d) select appropriate values for the training parameters and perform the training of the H-VSM model; and (e) design an algorithm to classify related terms into groups of lexical relations. The below sections

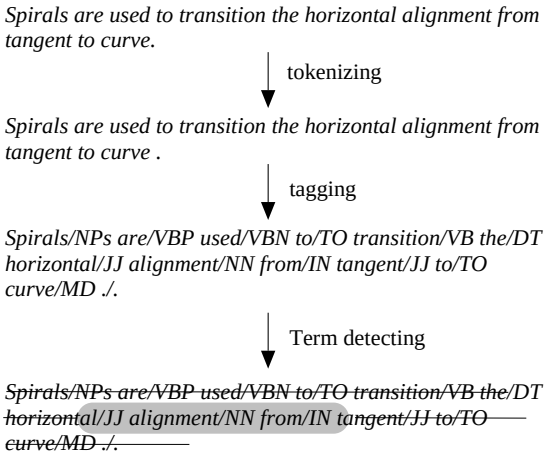

FIG. 2: Linguistic processing procedure to detect technical terms

discuss these steps in detail.

Data collection

As mentioned earlier, H-VSM was trained using a machine learning model which requires a text corpus as the source of the training dataset. The input text corpus was built upon a plethora of highway engineering manuals from the Federal Department of Transportation (DOT) and from 22 State DOTs. These documents in American-English. The technical terms in a guidance document in the engineering field are organized in various formats such as plain text, tables, and equations. Since words in tables and equations are not yet supported by the state-of-the-art NLP techniques, they were removed from the text corpus. The removal of these features would reduce corpus size, but it is necessary since words in tables and equations are not organized in a regular sentence structure and therefore the NLP algorithm may extract unreal noun phrases. The final outcome of this phase is a plain text corpus consisting of 16 million words. This dataset is utilized to extract highway related technical terms which are then trained and converted into representation vectors.

Multi-word terms extraction

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

terms may not be directly interpreted from the meanings of their single words. In order for the Skip-gram model to learn the semantics of multi-word terms, every occurrence of multi-word terms in the corpus needs to be detected and replaced with connected blocks of word members so that they can be treated as single words. This research utilizes OpenNLP, NLP package, to process the collected corpus and detect sequences of words that match pre-defined noun phrase patterns. Figure 2 presents the process of detecting technical terms from the set of highway technical documents. The process includes the following steps.

i **Word tokenizing:** In this step, the text corpus is broken down into individual units (also called tokens) using OpenNLP Tokenizer.

ii **Part of Speed (POS) tagging:** The purpose of this step is to determine the Part of Speech (POS) tag (e.g., noun, adjective, verb, etc.) for each token of the tokenized corpus obtained from the previous step. A set of POS tags can be found in the Penn Treebank (Marcus et al. 1993).

iii **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 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 Justeson and Katz (1995) to extract NPs. The tagged corpus is thoroughly scanned, and sequences matching to the noun phrase patterns is collected. In addition, in order to avoid discrimination among the syntactic variants of the same term, for example 'roadway' and 'roadways', the collected NPs need to be normalized. The following are two 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 allow for automated removal of suffixes,

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
Note: |, * respectively denote 'and/or', and 'zero or more'.	

is applied on the extracted noun phrases to normalize plural nouns (NNS) into single nouns (NN). Since the stemming algorithm affects only on the NNS token of a Noun phrase, the issue of over and under stemming can be minimized/eliminated.

- **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 is converted into the non-preposition form by removing the preposition and reversing the order of the remaining portions. For instance, 'type of roadway' will become 'roadway type'.

Since NPs with low occurrence frequencies that are unlikely to be technical terms should be automatically eliminated. With the frequency threshold of 2, the list consists of 112,024 terms. The list size drops to 8,922 when a threshold of 50 is used. In this research we used a threshold of 50.

iv **Multi-word term candidate ranking and selection:** Multi-word term definition varies between authors, and there is a lack of formal and widely accepted rules to define if a NP is a multi-word term (Frantzi et al. 2000). There are a number of methods proposed for estimating 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 the 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 the 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, for instance studies performed by Ananiadou et al. (2000), Lossio-Ventura et al. (2013), and Nenadić et al. (2002). Since the corpus used in this study was mainly collected from technical domain documents, C-value would be the most suitable for the termhood determination task. The C-value measure, as formulated in Equation 1, suggests that the longer a NP 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

|a| is the length of noun phrase a

f is the frequency of a in the corpus

Ta is the set of extracted noun phrases that contain a

P(Ta) is the size of Ta set.

The term extraction process above results in a dataset containing the detected terms along with their c-value termhood scores. These term candidates are ranked by C-value, and the ones that have negative C-values are discarded.

To remove candidates that are unlikely to be real terms, a threshold C-value can be used or the entire candidate list should be manually evaluated by industry experts. Manual evaluation would avoid the removal of real terms but have low C-values. To minimize both laborious work and the number of true terms wrongly discarded, the ranked list of candidate

TABLE 2: Excerpts of the extracted candidate terms

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

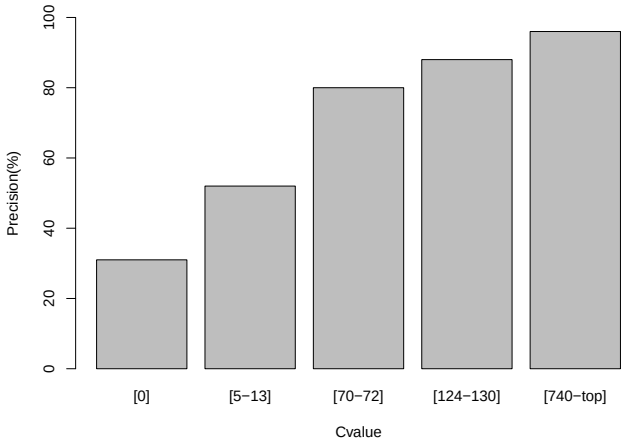

FIG. 3: Multi-word term extraction evaluation

were divided into groups of 100 items. A graduate student with civil engineering background was asked to utilize a bottom-up approach to evaluate group by group and stop at which has a precision of 80 percent. Table 2 illustrates the evaluation results for several excerpts of the extracted term candidates. The precision values, which represent the percentage of real terms in these groups, are presented in Figure 3. As shown in the figure, precision values are relatively low for groups with c-values less than 70. To balance between human effort and precision of the final term list, this research applied a manual review on the set of XX automatically extracted terms with c-values less than the threshold of 70 at the bottom of the list.

Construction of term space model

This step aims at processing the collected text corpus and collecting the training data for developing the H-VSM model. Skip-gram (Mikolov et al. 2013), which is an un-supervised

machine model, was employed to learn the semantic similarity among words in the text corpus. 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 which are closed to the input unit in the corpus. In order to collect this training dataset, the unannotated highway corpus is scanned to capture instances of terms and their corresponding context words. Each occurrence of a word 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 fulfill that requirement, every occurrence of a certain multi-word term in the corpus is replaced with a single unit that is compiled by connecting all the individual words. For instance, 'vertical alignment' becomes 'vertical-alignment'.

The number of context words to be collected is dependent on the window size that limits how many words to the left and the right of the target word. In the example sentence below, the context of term 'roadway' with the window size of 5 will be the following word set {bike, lane, width, on, a, width, no, curb, gutter}. Any context word that is in the stop list (the list contains frequent words in English such as 'a', 'an', and 'the' that have little meaning) will be neglected from the context set.

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

The semantic similarity is trained using the Word2vec module in the Apache Spark MLlib package (Apache.org 2016), an emerging open-source engine, which is based on the Skip-gram neural network model (Mikolov et al. 2013). Figure 4 shows the learning network when the context set includes only one word, where V and N respectively denote the corpus vocabulary and hidden layer size. In this model, a word in the corpus vocabulary is encoded as a 'one-hot' vector which is a vector in which only one elements at the index of the word in the vocabulary is set one, and all other items are zero. For example, the one-hot vector

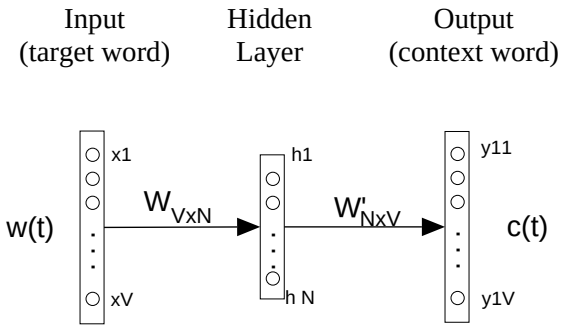

FIG. 4: Skip-gram model

of k^{th} word in the vocabulary with the size of V will be x1=0, x2=0, ...,xk=1,...xV=0. The outcome of this machine learning process is a set of term representation vectors in an N-dimension coordinate system. as we can see, the similarity among predicted context vectors are decided by the similarity of the corresponding *representation vectors*. each row of the W matrix which is the output of the learning process, is a representation vector of a word in the corpus vocabulary. The similarity among these vectors represent the similarity of the context of the corresponding words.

1. k^{th} input word : $[x_k]_{1.V} = [x_1 = 0, x_2 = 0, ..., x_k = 1, ..., x_V = 0]$ which is an one-hot vector.

2. Hidden vector:$[h]_{1.N} = [x_k]_{1.V}.W_{V.N} = [w_{k1}, w_{k2}, ..., w_{kN}] = v_{wk}$ which is equivalent to the k^{th} row of the W matrix since the input vector is a 'one-hot' vector. The v_{wk} vector is called the input *representation vector* of the input word k^{th}.

3. Predicted context vector: $[y_k]_{1.V} = v_{wk}.W'_{N.V}$.

The model includes three major parameters that are *frequency threshold*, *hidden layer size* and *window size* (see Table 3). To eliminate those data points with low frequencies of occurrence that are unlikely to be technical terms, Word2vec allows for the use of *frequency threshold*. Any word with the rate lower than the limit will be ignored. Radim (2014) suggests a range of (0-100) depending on the data set size. Setting this parameter high will enhance the accuracy, but many true technical terms would be out of vocabulary. A preliminary

TABLE 3: Skip-gram model parameters

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

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. The reasonable values for this parameter are from ten to hundreds (Radim 2014). The final major parameter, *context window size*, decides how many context words to be considered. Google recommends the size of 10 for the Skip-gram model (Google Inc. 2016). These parameters are subject to be changed so that the best model can be achieved. The effects of these parameters on the model performance are discussed in Section 4.

Figure 5 presents the term space model of H-VSM derived from the training process when the parameters are set 50, 300 and 10 respectively. H-VSM currently consists of more than 6,000 technical terms. In this model, each technical term is represented as a vector in a high dimensional space. Since the term representation vectors are in a multi-dimensional space; to present the space in 2D graph, PCA (Principle Component Analysis) was used to reduce the size to 2.

The similarity between terms in the H-VSM model can be measured by the angle between two word representation vectors (Equation 2) or the distance between two word points (Equation 3). Figure 5 illustrates the clustering of terms by their distances. In this figure, an *inlet* can be inferred to be more similar to an *outlet* (blue) than a *sidewalk* (green). Using this technique, the most similar terms for a given term can be obtained. Table 4 shows a

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

partial ranked list of the nearest terms of 'roadway' in order of similarity score.

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

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

Where: n is the hidden layer size.

Construction of term lexical hierarchy

The purpose of this module is to construct Infralex, a lexicon of civil engineering technical terms. A lexicon, also known as a lightweight knowledge base, typically includes terms and relations. The core relations of a lexicon are 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). Two terms that relate

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

each other through these semantic relations would have a high similarity score. 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 'roadway'. In this list, the true synonyms are 'highway' (1), 'traveled-way' (2) and 'road' (4); the attributes include 'roadway-section' (3), 'roadway-shoulder' (12); and 'adjacent-roadway' (7) and 'undivided' (37) are hyponyms which show different types of roadway.

The specific objective of this task is to detect the semantic relations among terms which are used for rearranging the nearest terms obtained from the H-VSM model. Algorithm 1 shows the design pseudo code for classifying the nearest terms of a given target term. The algorithm utilizes linguistic rules and clustering analysis to organize the nearest list into the following three groups: (1) attribute, (2) hyponym, and (3) synonym/sibling. The algorithm, first detects terms belonging to the first two categories using linguistic patterns. The filter rules to detect these relations are presented in Table 5. For a multi-word term matching pattern 1, we can infer that *Noun1* is an attribute of concept *Noun2*; and *Noun2* is an

attribute of *Noun1* in the pattern 2. Pattern 3 is for detecting hyponyms where the matched NP is a hyponym of *Noun2* concept. The remained nearest 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 the similarity score. The terms in the last group are unlikely to be a synonym or sibling; and thus, are removed from the nearest list. The output of the proposed algorithm is a list of classified nearest terms. 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
```

PERFORMANCE EVALUATION

This section presents a performance evaluation of InfraLex on the ability to identify synonyms. In this experiment, a gold standard is used. The gold standard consists of 70 sets of synonyms (both single and multi-word terms) which were examined and extracted

TABLE 5: Patterns to extract attributes and hyponyms

Relation	**Pattern**	**Example**
Attribute	Noun1 of Noun2	the width of the road
	Noun1 Noun2	road width, project cost
Hypernym-hyponym	Noun1 Noun2	vertical alignment isA alignment

TABLE 6: An example in InfraLex

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

from a Wikipedia transportation glossary (Wikipedia 2016). The developed Infralex model was employed to find the synonym for a given input term. The automatically identified synonym is the nearest word in the synonym/sibling lexical group. The evaluation outcome returns "true" if the automatically identified synonym belongs to the actual synonym set of the tested term in the golden standard. The performance was evaluated using the following three measures including precision, recall, and f-measure. Precision refers the accuracy in the conclusions made by the system, and recall reflects the coverage of domain terms of the system. The F score, which is a combined measure of precision and recall, presents the

overall performance of a system.

$$Precision = \frac{\text{number of correctly detected synonyms}}{\text{total detected terms}} \tag{4}$$

$$Recall = \frac{\text{number of correctly detected synonyms}}{\text{total terms}} \tag{5}$$

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

Table 7 shows the performance with various training model settings. The parameters of the baseline model are 50, 100 and 5 respectively for frequency threshold, hidden layer and window size. The authors changed these parameters one by one and kept the other ones unchanged 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 has a precision of 81% and an F-measure of 65%. The change of other parameters did not improve the performance. Especially, the increase of frequency threshold value has negative impact.

The proposed model was also compared with the generic Wordnet database. Table 8 presents the comparison of performance between InfraLex (with the 50-100-10 setting) and

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**

Wordnet. As shown, InfraLex outperforms Wordnet in all measures, and the combined F-measure is significantly improved (65% compared to 52%). The biggest contribution to the improvement of the overall F-measure is the recall value which represents a better coverage of domain vocabulary of InfraLex.

DISCUSSIONS

the proposed method to construct lexicon for the construction industry enable a quick translation of text documents to lexicon. The application of the method for roadway domain, a large dataset of terminology with more than 6000 terms have been quickly capture and the relations between terms are be able extracted as well. the research is expected to leverage the and scale up the whole infrastructure level to developed a comprehensive machine readable dictionary for the industry which data integrating and sharing systems can eliminate any terminology mismatches when integrating data from multiple sources. The lexicon dataset developed in this study is expected to become a fundamental resource for a variety of NLP related studies in the civil infrastructure domains. InfraLex can serve as a machine-readable dictionary of domain technical terms. NLP based platforms can utilize this resource for term sense analysis which is crucial for text mining to extract meaningful information from text documents, information retrieval, or natural language based human-machine interaction. Some specific examples of these potential applications are as follows. First, information retrieval systems can use the semantic relations provided by InfraLex to classify project documents by relevant topics by analyzing the keywords in the documents. Second, questionnaire designers can utilize InfraLex to search for synonyms so that appropriate terms can be selected for specific groups of potential respondents who might be from multiple disciplines or regions. Another application is that the query systems for extracting data from 3D engineered models would be able to find alternative ways to query data when users' keywords do not match any entity in the database. Since users have different ways and keywords to query data, the ability to recognize synonyms and related concepts of a query system would provide flexibility to the end user. Also, the developed InfraLex lexicon would

enable the matching data items such as (e.g., cost, productivity, etc.) when integrating data from distinct departments or states to develop a national database. This study is also expected to fundamentally transform the way human interacts with machine as technical terms which are a basic unit of human language can be precisely understood by computer systems. Instead of using computer languages, the end user can use natural language to communicate with computer systems. In order to enable computer to understand human language, a machine-readable dictionary which defines meanings of relevant vocabulary is required. therefore, the developed lexicon can be used by NLP applications for the domain of infrastructure.

The current study has some limitations that may contribute the low overall performance. First, the highway corpus is still relatively small with only 16 million words, compared to the corpus sizes in other domains with billions of words. Since the recall value largely depends on the corpus size, the expansion of the highway corpus would enable more technical terms to be covered in InfraLex. Future research is needed to enhance the performance of InfraLex by enlarging the data training set in both size and the number of disciplines involved throughout the life cycle of a highway project, such as asset management, project programming, construction management. The corpus also needs to cover other types of transportation assets like bridge, tunnel, railway, culvert, etc. Another work that can potentially improve the model performance is to distinguish synonym and sibling which are still in the same group in the InfraLex system. When these two lexical relations are separated, the possibility of recognizing a wrong synonym will be reduced; and consequently, the precision value would be enhanced.

CONCLUSIONS

Data manipulation from multiple sources is a challenging task in infrastructure management due to the inconsistency of data format and terminology. The contribution of this study is a digital lexicon of highway related technical terms (named InfraLex) which can enable a computer to understand semantic meanings of terms. This research employs advanced

NLP techniques to extract technical terms from a highway text corpus which is composed of 16 million words built on a collection of design manuals from 22 State DOTs across the U.S. Machine learning was used to train the semantic similarity between technical terms. An algorithm was designed to classify the nearest terms resulted from the semantic similarity model into distinct groups according to their lexical relationships. This algorithm was employed to develop the InfraLex database.

The developed lexicon has been evaluated by comparing the results obtained from the computational model and a man-crafted gold standard. The result shows an accuracy of over 80 percent. The best model is associated with the training parameters of 50, 100 and 10 respectively for frequency threshold, hidden layer size, and window size. Although significant improvement is shown in comparison with the existing thesaurus databases, the overall performance is not relatively high. This might be due to the size of the training data. Future research will be conducted to expand the highway corpus to further disciplines such as asset management, and transportation operation.

The research opens a new gate for computational tools regarding natural language processing in the highway sector. InfraLex would enable computer systems to understand terms and consequently transform the way human interacts with computer by allowing users to use natural language.

REFERENCES

(2008). "Building and civil engineering vocabulary - part 4: Transportation.

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).

627 buildingSMART (2016). "Data dictionary, <http://www.buildingsmart.org/standards/standards-
628 library-tools-services/data-dictionary/>. Accessed: March 15, 2016.
629 Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., and Racioppa, S. (2008). "Ontology-based
630 information extraction and integration from heterogeneous data sources." *International*
631 *Journal of Human-Computer Studies*, 66(11), 759–788.
632 Cai, H., Yuan, C., McClure, T. B., and Dunston, P. S. (2015). "A synthesis study on collect-
633 ing, managing, and sharing road construction asset data.
634 Cambria, E. and White, B. (2014). "Jumping nlp curves: a review of natural language
635 processing research [review article]." *Computational Intelligence Magazine, IEEE*, 9(2),
636 48–57.
637 Chen, D. and Manning, C. D. (2014). "A fast and accurate dependency parser using neural
638 networks." *EMNLP*, 740–750.
639 Church, K. W. and Hanks, P. (1990). "Word association norms, mutual information, and
640 lexicography." *Computational linguistics*, 16(1), 22–29.
641 Costa-Jussa, M. R., Farrús, M., Mariño, J. B., and Fonollosa, J. A. (2012). "Study and
642 comparison of rule-based and statistical catalan-spanish machine translation systems."
643 *Computing and Informatics*, 31(2), 245–270.
644 Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). "Gate: an architecture
645 for development of robust hlt applications." *Proceedings of the 40th annual meeting on*
646 *association for computational linguistics*, Association for Computational Linguistics, 168–
647 175.
648 El-Diraby, T. and Kashif, K. (2005). "Distributed ontology architecture for knowledge man-
649 agement in highway construction." *Journal of Construction Engineering and Management*,
650 131(5), 591–603.
651 El-Diraby, T., Lima, C., and Feis, B. (2005). "Domain taxonomy for construction concepts:
652 Toward a formal ontology for construction knowledge." *Journal of Computing in Civil*
653 *Engineering*, 19(4), 394–406.

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.

Gallaher, M. P., O'Connor, A. C., Dettbarn, J. L., and Gilday, L. T. (2004). *Cost analysis of inadequate interoperability in the U.S. capital facilities industry*. U.S. Department of Commerce Technology Administration, National Institute of Standards and Technology, Gaithersburg, MD.

Google Inc. (2016). "word2vec, <https://code.google.com/archive/p/word2vec/>. (accessed May 12, 2016).

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*.

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

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

Hill, F., Reichart, R., and Korhonen, A. (2015). "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." *Computational Linguistics*, 41(4), 665–695.

Hsieh, S.-H., Lin, H.-T., Chi, N.-W., Chou, K.-W., and Lin, K.-Y. (2011). "Enabling the development of base domain ontology through extraction of knowledge from engineering domain handbooks." *Advanced Engineering Informatics*, 25(2), 288–296.

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

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.

Karimi, H. A., Akinci, B., Boukamp, F., and Peachavanish, R. (2003). "Semantic interoperability in infrastructure systems." *Information Technology*, 42–42.

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

Landauer, T. K. and Dumais, S. T. (1997). "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.." *Psychological review*, 104(2), 211.

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

Lefler, N. X. (2014). "Roadway safety data interoperability between local and state agencies." *Report no.*

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

Levy, O., Goldberg, Y., and Dagan, I. (2015). "Improving distributional similarity with lessons learned from word embeddings." *Transactions of the Association for Computational Linguistics*, 3, 211–225.

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

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

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2013). "Combining c-value and keyword extraction methods for biomedical terms extraction." *LBM'2013: 5th International Symposium on Languages in Biology and Medicine*.

Lv, X. and El-Gohary, N. M. (2015). "Semantic annotation for context-aware information retrieval for supporting the environmental review of transportation projects." *Computing in Civil Engineering 2015*, ASCE, 165–172.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA., 281–297.

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

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). "Building a large annotated corpus of english: The penn treebank." *Computational linguistics*, 19(2), 313–330.

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.

Mounce, S., Brewster, C., Ashley, R., and Hurley, L. (2010). "Knowledge management for more sustainable water systems." *Journal of information technology in construction*, 15, 140–148.

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

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

Niu, J. and Issa, R. R. A. (2015). "Developing taxonomy for the domain ontology of construction contractual semantics: A case study on the aia a201 document." *Advanced Engineering Informatics*, 29(3), 472–482.

Noy, N. F. (2004). "Semantic integration: a survey of ontology-based approaches." *ACM Sigmod Record*, 33(4), 65–70.

735 Osman, H. and Ei-Diraby, T. (2006). "Ontological modeling of infrastructure products and
736 related concepts." *Transportation Research Record: Journal of the Transportation Research*
737 *Board*, 1984(-1), 159–167.
738 Ouksel, A. M. and Sheth, A. (1999). "Semantic interoperability in global information sys-
739 tems." *ACM Sigmod Record*, 28(1), 5–12.
740 Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: Global vectors for word
741 representation." *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543,
742 <http://www.aclweb.org/anthology/D14-1162>.
743 Porter, M. F. (1980). "An algorithm for suffix stripping." *Program*, 14(3), 130–137.
744 Radim, R. (2014). "Word2vec tutorial, <http://rare-technologies.com/word2vec-tutorial/>.
745 Rezgui, Y. (2007). "Text-based domain ontology building using tf-idf and metric clusters
746 techniques." *The Knowledge Engineering Review*, 22(04), 379–403.
747 Salton, G. and Buckley, C. (1988). "Term-weighting approaches in automatic text retrieval."
748 *Information processing & management*, 24(5), 513–523.
749 Sclano, F. and Velardi, P. (2007). *Termextractor: a web application to learn the shared*
750 *terminology of emergent web communities*. Springer, 287–290.
751 Seedah, D. P., Choubassi, C., and Leite, F. (2015a). "Ontology for querying heteroge-
752 neous data sources in freight transportation." *Journal of Computing in Civil Engineering*,
753 04015069.
754 Seedah, D. P., Sankaran, B., and O?Brien, W. J. (2015b). "Approach to classifying freight
755 data elements across multiple data sources." *Transportation Research Record: Journal of*
756 *the Transportation Research Board*, (2529), 56–65.
757 Seng, J.-L. and Kong, I. L. (2009). "A schema and ontology-aided intelligent information
758 integration." *Expert Systems with Applications*, 36(7), 10538–10550.
759 Sparck Jones, K. (1972). "A statistical interpretation of term specificity and its application
760 in retrieval." *Journal of documentation*, 28(1), 11–21.
761 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.

Turney, P. D. and Pantel, P. (2010). "From frequency to meaning: Vector space models of semantics." *Journal of artificial intelligence research*, 37(1), 141–188.

Walton, C. M., Seedah, D. P., Choubassi, C., Wu, H., Ehlert, A., Harrison, R., Loftus-Otway, L., Harvey, J., Meyer, J., Calhoun, J., et al. (2015). *Implementing the freight transportation data architecture: Data element dictionary*. Number Project NCFRP-47.

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.

Wetherill, M., Rezgui, Y., Lima, C., and Zarli, A. (2002). "Knowledge management for the construction industry: the e-cognos project.

Wikipedia (2016). "Glossary of road transportation terms. Accessed: April 11, 2016.

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." *5th International/11th Construction Specialty Conference*.

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.