InfraLex: An automatically-generated infrastructure lexicon for matching data entities and attributes from heterogeneous sources

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

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

ABSTRACT

The inconsistency of data terminology due to the fragmented nature of the civil industry has imposed big challenges for integrating life-cycle digital data from distinct sources to support decision making in highway asset management. The issue of data ambiguity may lead to the lack of common understanding to the same data between the data sender and receiver. While the aspect of data structure has been well addressed thanks to the availability of various international neutral data standards such as LandXML and TransXML; the semantic aspect still has been neglected by researchers. This paper presents a novel methodology to construct an automatically-generated lexicon which is consisted of commonly used technical terms for civil infrastructure projects. The lexicon provides an underlying resource for computational systems that perform semantics analysis in the civil infrastructure domain. 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 U.S. 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 to organize related terms into separate groups according to their semantic relations. The developed lexicon has been experimented on the ability of recognizing the most semantically similar terms and achieved a precision of 80 percent.

[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.

Keywords: Civil infrastructure project, Lexicon, Data retrieval, Natural language interface, 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, translates 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 models as an accessible, reusable and reliable information source for life-cycle decision making. According to a study conducted by the National Institute of Standard and Technology (NIST), the un-interoperability issue was reported to cost the U.S. capital facilities industry at least $15.8 billion per year, and two-thirds of those costs were incurred during the operation and maintenance stages (Gallaher et al. 2004). The major cost was time spent 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.

Due to the fragmented nature of the construction industry, data terminology varies between phases, stakeholders, or geographic regions (counties, states, etc.). Polysemy and synonymy are two major semantic obstacles to the integration and use of data from multiple sources. Polysemy refers to cases when a unique term has distinct meanings. For example, the term 'roadway type', in a generic 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. In contrast, synonymy is associated with the diversity of terms 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 or improper extraction of the desired data. Thus, addressing the terminology ambiguity issue becomes critical 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.

Research to address the issue of terminology inconsistency in the construction industry is limited. Due to this reason, although an extensive amount of research effort has been made for last several decades in standardizing a neutral data format, such as Industry Foundation Classes (IFC) (buildingSMART 2015) or LandXML (landxml.org 2015), the vision of seamless exchange of data is still in slow progress. One of the approach to enable proper and ambiguity-free reuse of digital data is to develop Model View Definitions (MVD) (buildingSMART 2016b) which formally identifies a mapping matrix between the data entities in a neutral data format and the local domain data labels for a specific transaction scenario. MVD enables partial models to be easily extracted and unambiguously interpreted. However, this approach is on a case-by-case and manual basis, and takes years to be completed and maintained; there is demand for more rigorous computational techniques that can allow for automated extraction of data with minimized human interfere (Venugopal et al. 2012; Eastman 2012). In addition to MVD, a few construction domain specific semantic resources have been proposed, for example the Civil Engineering Thesaurus (CET) (Abuzir and Abuzir 2002), e-Cognos (Wetherill et al. 2002), and buildingSMART data dictionary (ISO 12006-3) (buildingSMART 2016a). These knowledge bases present domain concepts in machine-readable format and allow computer systems to understand meanings of terms. Thus, data mismatch when unifying isolated data sources would be eliminated. They, however, are mainly hand-coded. Due to this reason, the existing domain digital dictionaries still cover only a small portion of the civil infrastructure related concepts. Therefore, there is a demand for computational techniques that can automatically construct and maintain these digital dictionaries to keep up with the increasingly arising of new terms.

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 for single linguistic units such as Part of Speech (POS) tagging (Toutanova et al. 2003; Cunningham et al. 2002), to relationships between linguistic units like dependency parser (Chen and Manning 2014). These basic NLP techniques have been applied in various computational methods that can support 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 considerable 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 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 process of translating text based domain knowledge into InfraLex, an extensive civil engineering machine-readable dictionary of technical terms. In order to achieve that goal, basic 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 U.S. 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 to organize related terms into separate groups according to their semantic relations. The framework was complied into a Java package and a lexicon dataset which can be found at https://github.com/tuyenbk/mvdgenerator.

The paper is organized as follows. This section presents background and rationale for the study. Section 2 provides underling knowledge supporting the study and gap of knowledge. Section 3 and 3 respectively describes the methodology employed to develop InfraLex and the performance evaluate results. Research limitations and potential applications will be discussed in Section 5. The final section concludes the paper with the discussion on major

findings and future research.

RELATED RESEARCH

This section will first present 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 collection of techniques that can analyze and extract information from natural languages like text and speech. The major applications of NLP include translation, information extraction, opinion mining (Cambria and White 2014). 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 individual 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 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 are therefore relatively low. NLP research is shifting to statistical ML-based methods (Cambria and White 2014). ML models are able to accurately learn patterns from training examples to predict the output. Hence, they are independent of languages and linguistic grammars (Costa-Jussa et al. 2012).

Vector space model

Measuring semantic similarity between semantic units (words, phrases, sentences, concepts, etc.) is one of the main NLP-related research topics. There are two major approaches for semantic measure including (1) dictionary-based method and (2) distributional method

(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). 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). A distributional model stands on the distributional hypothesis that states that two similar terms would occur in the same context (Harris 1954). The outcome of this approach is a Vector Space Model (VSM), 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). VSM outperforms the dictionary-based method in terms of time saving as the semantic model can be automatically obtained from text corpus and corpus collecting is much easier than manually constructing a digital dictionary (Turney and Pantel 2010). 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 a set of representation vectors for technical terms.

The VSM approach has been 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 have evidently demonstrated the promising of the VSM in identifying the semantic similarity between technical terms in order to develop advanced tools for handling data stored in natural language documents generated through the project life cycle.

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 a machine to perform knowledge works such as interpreting users' intention or understanding human-oriented inputs. However, there is still a shortage of such an extensive dictionary for the civil engineering domain. WordNet (Miller 1995), which is one of the largest lexicons with over 117,000 synsets for NLP related applications, 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), e-Cognos (Wetherill et al. 2002), and buildingSMART data dictionary (ISO 12006-3) (buildingSMART 2016a). 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). Like other construction specific digital dictionaries, buildingSMART dictionary is mainly hand-coded and time consuming; the vocabulary, therefore, is still relatively limited. 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 studies have been done to deal with the issue of sense ambiguity. (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 remains in cases where the same word form is used for different senses. Additionaly, ZESeM relies on Wordnet which lacks highway technical terms, NLP-based frameworks built upon this algorithm would have low performance. (Lin et al. 2015) developed an 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 in which synonyms are assigned 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 compared to the large collection of terms in the construction industry.

INFRALEX CONSTRUCTION

Overview of the proposed methodology

The ultimate goal of this research is to construct a machine-readable dictionary of technical terms, named InfraLex, for the infrastructure sector. This research proposes a methodology for automated construction of the domain thesaurus using advanced Natural NLP techniques.

Figure 1 presents an overview of the methodology proposed to develop InfraLex. The research framework is consisted of two major modules that are to: (1) train a highway term space model (H-VSM), and (2) develop an algorithm integrating H-VSM and linguistic

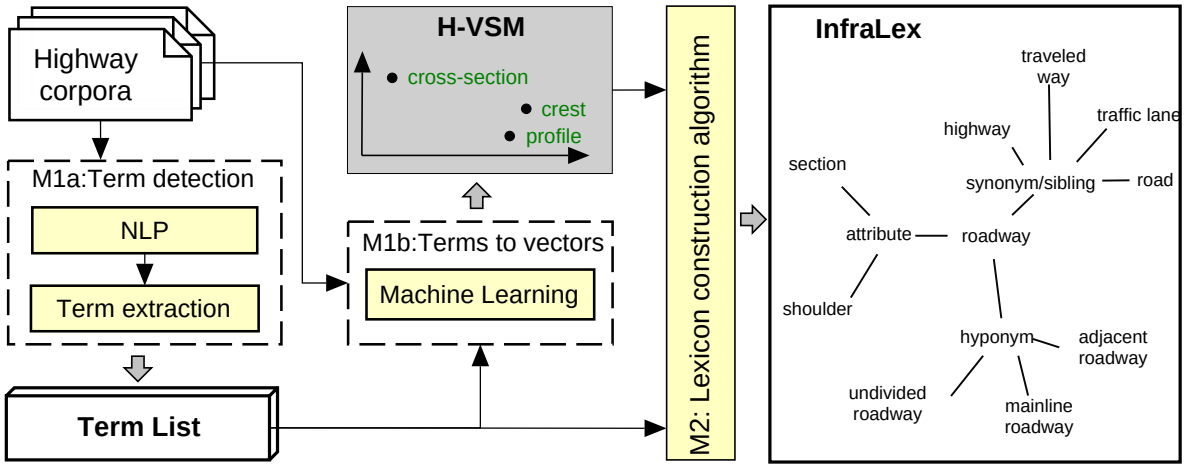

FIG. 1: Overview of the proposed methodology

patterns to construct InfraLex. The first module implements several basic NLP techniques (including tokenizing, POS tagging, etc.) and C-value (Frantzi et al. 2000) method to extract highway related technical terms from a highway corpora. Skip-gram model, an unsupervised machine learning platform proposed by (Mikolov et al. 2013), is then implemented 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 degree of similarity between technical terms can be determined; and based on that a list of the nearest terms 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 by semantic relations such as synonymy, sibling, hypernymy, hyponymy and attribute. Specifically, the procedure followed to compile the InfraLex dictionary is comprised of the following steps which will be discussed in detail in the below sections.

- Collect highway technical documents to compose a domain corpus;
- Extract multi-word terms from the highway corpus;
- Prepare training dataset for training H-VSM;
- Selecting training parameters and train H-VSM;

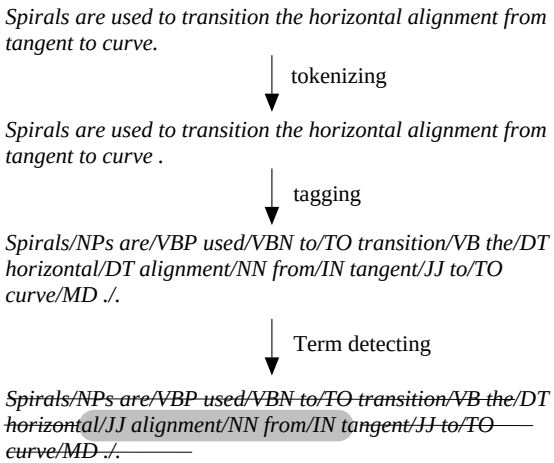

FIG. 2: Linguistic processing procedure to detect technical terms

- Design an algorithm to classify related terms into groups of lexical relations.

Data collection

As mentioned earlier, 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 22 distinct State DOTs. The focus of this corpora 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 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 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 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. Figure 2 presents the process of detecting technical terms from the set of highway technical documents. The process includes the following steps.

1. **Word tokenizing:** In this step, the text corpus is broken down into individual units (also called tokens) using OpenNLP Tokenizer.
2. **Part of Speed tagging:** The purpose of this step is to determine the POS tag (e.g., noun, adjective, verb, etc.) of 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 (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 written in conjunctions (e.g., 'vertical and horizontal alignment'). To extract NPs from a conjunction, an extra processing is applied to break it into individual NPs. 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 NPs and the dependent parts ('vertical' and 'horizontal'). This research uses Stanford Dependencies Parsing tool, which is able to analyze dependencies between sentiment units, to split conjunctions 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.

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'.	

- **Type 1** - Plural forms, for example 'roadways' and 'roadway'. The Porter stemming algorithm [(Porter 1980)], which can allow for 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 NP of an abbreviation. This method suggests the following abbreviation definition 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 is a lack of formal rules for defining multi-word term (Frantzi et al. 2000). There are a number of methods 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 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 study was 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 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

f is the frequency of a in the corpus

Ta is the set of extracted noun phrases that contains a

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

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

To remove non-terms from the term list, a manual evaluation process was conducted. Table 2 shows the evaluation result for 5 examples of the extracted terms. The longer the list is, the more effort required for the evaluation process. Since 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 112,024 terms. The list size drops to 8,922 when a

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

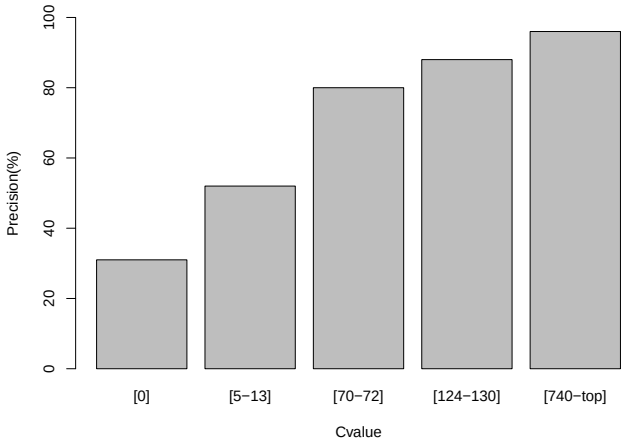

FIG. 3: Precision of term extraction

threshold of 50 is used. Manually reviewing such a long list is still a challenging task. To minimize human force, the list was evaluated at several ranges of C-values. Precision, which represents the percentage of real terms in each group, is presented in Figure 3. As shown in the figure, precision values are relative low for groups with c-values less than 70. To balance between human cost and precision of the final term list, this research applied the manual review on all of the automatically extracted terms below the c-value threshold of 70.

Training dataset preparation

The highway text corpora 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

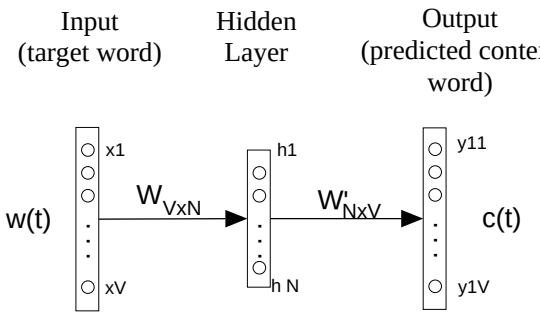

FIG. 4: Skip-gram model

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 multi-word term in the corpus is replaced with a single unit that is compiled by connecting all 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 the term 'roadway' with the context window size of 10 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', 'the' that have little meaning) will be neglected in the context set.

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

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). The model includes three major parameters that are frequency threshold, hidden layer size and window size (see Table 3). To eliminate data points with low frequency of occurrence that are unlikely to be technical

TABLE 3: Skip-gram model parameters

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

terms, word2vec includes the parameter of *frequency threshold*. 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. A 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. 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 is suggested to be in the range of 100-500. The final major parameter is *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 on the model performance are discussed in Section 4.

Figure 5 presents the term space model developed from the training process when the parameters are set 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 terms represents semantic similarity. H-VSM is consisted of more than 6,000 technical keywords. Since the vector space is a multi-dimensional space in accordance with the size of the hidden layer. In order to present the space in 2D graph, PCA (Principle

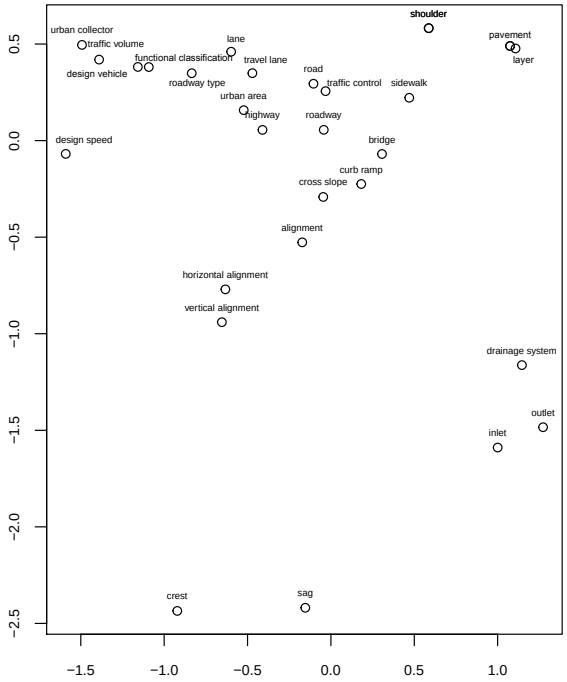

FIG. 5: Highway term space model (H-VSM)

Component Analysis) was used to reduce the size to 2 dimensions.

The similarity between terms can be measured by the angle between two word representation vectors (Equation 2) or the distance between two word points (Equation 3). Table 4 shows an example of a ranked list of near terms obtained from the H-VSM model 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.

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 a lexicon can be classified into the following types: synonym (meaning equivalence), hypernym-

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

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 relate each other through these semantic relation 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 the term 'roadway'. In this list, true synonyms are highway (1), traveled-way (2) or road(4); attributes include roadway-section (3), roadway-shoulder (12); and adjacent-roadway (7) and 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 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 reorganize the nearest list into the following three groups: (1) attribute, (2) hyponym, and (3) synonym/sibling and other functional relations. The algorithm firstly detects terms for 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 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

TABLE 5: Patters 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: 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

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 is consisting 70 sets of synonyms (both single and multi-word terms) which were examined and extracted from a Wikipedia transportation glossary (Wikipedia 2016). This glossary provides plain text explanation for each term and their synonyms. The automatically identified synonym which is the top nearest word of the synonym/sibling lexical group was compared with the true synonym in the gold standard dataset. The results are evaluated 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 parameters of the baseline model are 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 parameters did not improve the performance. Especially, the increase of frequency threshold has negative impact.

Table 8 presents the comparison of performance between InfraLex (with 50-100-10 set-

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**

tings) and Wordnet. As shown, InfraLex outperformed Wordnet in all aspects, 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 current study has several limitations that may be the reason for the low overall performance. Firstly, the highway corpus is still relatively small with only 160 million words, compared to corpus size in other domain. 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. Another work that potentially improve the model performance is to distinguish synonym and sibling which still in the same group in the proposed InfraLex. When these two lexical relations are separated, the possibility to recognize a wrong synonym will be reduced; and consequently, the precision value would be enhanced.

The lexicon dataset developed in this study is expected to become an underlying resource for a variety of NLP related studies in the civil infrastructure domains. InfraLex would serve as a machine-readable dictionary of domain technical terms. NLP based platforms can utilize this resource for word 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. Firstly information retrieval systems can use the semantic relations provided by InfraLex to classify project documents by topic. Secondly, questionnaire designers can utilize InfraLex to search for synonyms so that appropriate terms can be selected in accordance with each group of potential respondents which may be from multiple disciplines or regions. Another application is that query systems to extract 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 would enable the matching data items such as (e.g., cost, productivity, etc.) when integrating data from distinct departments or states. Last but not least, this study is expected to fundamentally transform the way human interacts with machine. Instead of using computer languages, the end user can use natural language to communicate with computer systems.

CONCLUSIONS

Data manipulation and retrieval from multiple sources is a challenging task 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 consisted 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 resulted 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 relative low. This may 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 pro-

cessing in the highway sectors. 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

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).

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

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

buildingSMART (2016b). "Model view definition summary, <http://www.buildingsmart-tech.org/specifications/ifc-view-definition>. (accessed April 12, 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.

Chen, D. and Manning, C. D. (2014). "A fast and accurate dependency parser using neural networks." *EMNLP*, 740–750.

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.

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.

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.

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*.

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.

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.

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.

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.

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*, http://lbm2013.biopathway.org/. Computer Science [cs]/Bioinformatics [q-bio.QM] Life Sciences [q-bio]/Quantitative Methods [q-bio.QM] Computer Science [cs]/Document and Text ProcessingConference papers.

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.

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

548 Miller, G. A. (1995). "Wordnet: a lexical database for english." *Communications of the ACM*,
549 38(11), 39–41.

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

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

555 Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: Global vectors for word
556 representation." *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543,
557 <http://www.aclweb.org/anthology/D14-1162>.

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

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

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

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

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

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

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

574 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.